DOT/FAA/TC-16/38

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

Separation Management Human-in-the-Loop Simulation: Evaluations of Conflict Probe Alert Reliability Requirements and Conflict Probe Algorithmic Improvements

Ben Willems, FAA Human Factors Branch, ANG-E25
Anthony J. Masalonis, Ph.D., Spectrum Software Technology, Inc.
Thomas Fincannon, Ph.D., Applied Research Associates, Inc.
George M. Puzen, Spectrum Software Technology, Inc.
Robert A. Bastholm, Spectrum Software Technology, Inc.

August 2016

Technical Report

This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

Technical Report Documentation Page

1. Report No. DOT/FAA/TC-16/38	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Separation Management Human-in-the-Loop Simulation: Evaluations of Conflict Probe Alert Reliability Requirements and Conflict Probe Algorithmic Improvements		**5. Report Date** August 2016 **6. Performing Organization Code** ANG-E25
7. Author(s) Ben Willems, FAA Human Factors Branch, ANG-E25 Anthony J. Masalonis, Ph.D., Spectrum Software Technology, Inc. Thomas Fincannon, Ph.D., Applied Research Associates, Inc. George M. Puzen, Spectrum Software Technology, Inc. Robert A. Bastholm, Spectrum Software Technology, Inc.		**8. Performing Organization Report No.** DOT/FAA/TC-16/38
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 Air Traffic Organization – Domain Engineering & Requirements Team 800 Independence Avenue, S.W. Washington, DC 20591		**13. Type of Report and Period Covered** Technical Report **14. Sponsoring Agency Code** AJV-731
15. Supplementary Notes		

16. Abstract

Objective: The purpose of this study is to assess the impact of factors that influence Conflict Probe (CP) Alert reliability (i.e., conformance bound settings, algorithmic enhancements, and radar surveillance noise) and display location (RA-Side vs. R- and RA-Side) on controller performance, attitude, and behavior. **Background**: This is the third in a series of studies examining how different automation and display characteristics impact a controller's ability to maintain separation between aircraft. The CP has emerged as an important tool to maintain separation. The current study examines how CP Alert reliability and display location impact a controller's ability to manage conflicts in a busy airspace. **Method**: Eight air traffic controllers participated in the experiment. The participants completed 12 sessions to examine the impact of algorithmic enhancements, radar surveillance noise, conformance bound and likelihood settings that determine which potential conflicts generate alerts at what timing, and CP display location. Because of technical limitations, conformance bound and likelihood settings did not vary across any of the conditions. **Results**: In spite of the aforementioned technical problems, which caused controllers to use a CP with a low level of alert reliability, controllers used the R-Side CP displays, and we identified a combined advantage of displaying the CP on the R-Side with algorithmic enhancements. The analysis of performance and subjective ratings indicated (a) that controllers only used the CP when it was on the R-Side; (b) that when the CP was available on the R-Side, it did improve controller performance; and (c) that controllers found the excess of yellow alerts in the Legacy (no algorithmic enhancements) condition to be a nuisance and a distraction from their tasks. **Conclusion**: Even at the lower levels of CP Alert reliability, which we inadvertently presented in more runs than intended, controllers accept CP functionality on the R-Side, and it provides some benefit. If R-Side CP is fielded with algorithmic enhancements and the current or better parameter settings, we expect acceptance and utility. Future research may need to consider the potential benefits of using a buffer zone to present near-conflicts to controllers. **Applications**: An R-Side CP that presents only red alerts will likely enhance controller performance, and controllers will likely accept it. Yellow alerts may represent information overload, but controllers indicate a need for some information about near-losses of separation. Future research may need to consider the potential benefits of using a buffer zone to present a limited number of near-conflicts.

17. Key Words Air Traffic Control Automation Reliability Conflict Probe Separation Management Workload		**18. Distribution Statement** This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov	
19. Security Classification (of this report) Unclassified	**20. Security Classification (of this page)** Unclassified	**21. No. of Pages** 175	**22. Price**
Form DOT F 1700.7 (8-72)	**Reproduction of completed page authorized**		

THIS PAGE IS BLANK INTENTIONALLY.

Table of Contents

List of Illustrations

Tables **Page**

PAGE IS BLANK INTENTIONALLY.

Acknowledgments

The authors would like to thank our sponsors, Mohamed Abouelenein from the Domain Engineering and Requirements Team (AJV-731), Wendy O'Connor from the ATO Operational Concepts Group (AJV-72), and Flora Emami from the ERAM Engineering Sub-team (AJM-2132) for their assistance and guidance during the planning, conduct, and analysis of this experiment. We are also grateful to the Air Traffic Control Specialists who participated in the study and to the controllers who served as observers.

The descriptions in this document of the experiment complications and solutions are taken from memos composed by several of the authors with major contributions from Kenneth Allendoerfer (ANG-E25) and Gary Mueller (ANG-E14). We also thank Gary and the DESIREE team for the extensive simulator development required to make this experiment happen.

PAGE IS BLANK INTENTIONALLY.

Executive Summary

Eight Certified Professional Controllers from the En Route Automation Modernization (ERAM) National User Team participated in a human-in-the-loop experiment to study the effects of three variables: (a) the availability of Conflict Probe (CP) data on only the Radar Associate (RA-Side) position versus on both the RA-Side and the Radar Side (R-Side) positions; (b) the presence of radar surveillance noise; and (c) the enhancement of CP algorithms on En Route Air Traffic Control performance, workload, and subjective opinions.

We conducted simulation runs where we placed CP information on the R-Side. We presented R-Side CP in the form of Full Data Block (FDB) portals and a Conflict Probe Alert View (CPAV) that contained a list of aircraft involved in predicted conflicts. The FDB safety portal indicated aircraft and airspace alerts. The safety portal presented CP Alerts in the form of a color-coded square, depicting the number of alerts for the aircraft and the color of the highest level alert. The CPAV depicted CP information in a tabular format with similar colored squares and minutes to go before conflict. In other runs, the CP information was available only on the RA-Side of the workstation. To manipulate the presence or absence of radar surveillance noise, our simulation software ran either with or without insertion of surveillance noise—absence of noise mimics a theoretical future situation in which surveillance is extremely accurate. To investigate the effect of CP algorithmic enhancements, we ran either a legacy version of the ERAM software or a software version that contained algorithmic enhancements that improved CP's conflict performance detection in engineering studies (Crowell, Fabian, Young, Musialek, & Paglione, 2011, 2012). In the condition where the enhancements were absent and the CP was on the R-Side, we presented yellow and muted alerts on both the R- and RA-Side displays. The yellow alert activates when CP predicts the conformance boxes surrounding the aircraft trajectories to come within minimum separation values. The red alert occurs when CP predicts the trajectory centerlines to violate separation. A muted alert occurs when the alert condition—red or yellow—will occur only if an aircraft proceeds on a "not yet cleared" portion of its trajectory, such as an assigned altitude. In other conditions with CP Alert information on the R-Side, red alerts and airspace alerts appeared on the R-Side. In all conditions, all alerts appeared on the RA-Side and were available in the Aircraft List and the Graphic Plan Display.

We assessed the effects of our manipulated variables on safety and efficiency and on objective and subjective workload. In the simulation runs, we presented traffic scenarios based on operational traffic recordings and modified to fit the needs of the experiment. All participants worked an airspace based on Sector 85 of the Indianapolis Air Route Traffic Control Center, which for this experiment, combined laterally coincident high-altitude, intermediate, and super high sectors resulting in an airspace consisting of all altitudes at and above Flight Level 240. We varied traffic levels over the course of each simulation run, increasing and decreasing traffic between 33% and 125% of the Monitor Alert Parameter. Each run took 50 minutes, and each participant ran 12 scenarios, plus make-up runs as needed, working the sectors alone over the course of three days preceded by a full day of training and practice.

The original experimental design also called for varying CP parameters shown to affect the CP Alert reliability of the conflict predictions (Crowell et al., 2011, 2012; Fincannon, Willems, & Masalonis, 2015), specifically lateral and longitudinal conformance bounds, and likelihood settings affecting R-Side Probe alert display timing. However, errors in the simulation setup prevented these

variables from changing between runs. This document focuses on the analysis of 6 of the 12 runs each participant conducted that ran mostly as intended.

Our objective data comprised safety- and efficiency-related data, such as losses of separation, frequency deviations, revisions to clearances, and meeting altitude restrictions. Other efficiency data included the time and distance that aircraft traveled in the sector; and objective indicators of controller activity such as voice communications, commands entered, and number of voice clearances as measured by simulation pilot commands. We also measured workload using functional Near-Infrared Spectroscopy (fNIRS) and subjectively via the Workload Assessment Keypad and participant questionnaires. These questionnaires, along with over-the-shoulder ratings, provided an additional source of performance data. Each participant ran all experimental conditions, so we conducted a repeated measures analysis to statistically test the effects of interest.

We predicted that CP information on the R-Side would enable controllers to resolve conflicts more proactively and generally improve performance and reduce workload. We predicted that controllers would indicate a favorable perception of locating CP on the R-Side. We hypothesized that enhancing CP Alert reliability by removing surveillance noise and implementing the algorithmic improvements would have the same effects.

Technical difficulties resulted in manipulation of the conformance bounds that was different from what we had designed. Conformance bounds were always set to 2.5 nautical miles (NM) (lateral) by 2.5 NM (longitudinal) with a 10-20 likelihood setting, but our Algorithmic Improvements and CP Location manipulations ran essentially as designed. In the Method and Discussion sections, we outline the lessons learned from these errors and the steps we have taken and will take to prevent them from repeating.

The fNIRS analysis suggested that algorithmic enhancements reduced controller workload. Controllers also reported fewer nuisance alerts and rated the CP as less of a distraction when the CP used algorithmic enhancements.

Integration of the CP Alerts into the R-Side display increased controller ratings of usefulness and over-the-shoulder observer ratings of participants' ability to (a) use the CP in appropriate situations, (b) use the CP in a timely and effective manner, and (c) use the CP overall. Controllers said that they were more likely to believe and respond to the R-Side CP, and an analysis of performance indicated that presenting a CP Alert on the R-Side more than halved the average time between issuing a suboptimal altitude clearance and addressing the problem with an amended clearance. Observer ratings did, however, suggest reduced situation awareness when CP was on the R-Side due to controllers reacting to the CP information instead of detecting conflicts on their own.

We found a few effects of noise (such as a possible reduction in workload as measured by fNIRS when noise was not present) and some effects on objective workload—but few, if any, effects that we deemed operationally meaningful.

Interaction effects between CP Location and CP Algorithm—for example, on questionnaire data—suggest that the CP should use algorithmic enhancements and should present red alerts, with no yellow alerts, on the R-Side display. However, a qualitative analysis indicated that controllers would be interested in a buffer zone, which enables alerting for near-losses of separation outside of the 5 NM threshold, and that their performance might benefit from it. We recommend further research specifically designed to address this topic and other topics arising from our results, such as trial planning and electronic controller-controller coordination.

Controllers were willing to accept a CP Alert reliability with levels below those of the current fielded system. This suggests that controllers are likely to accept integrated CP Alerts on R-Side.

PAGE IS BLANK INTENTIONALLY.

1. INTRODUCTION

1.1 Background

The Human Factors Branch (ANG-E25) conducted a human-in-the-loop (HITL) experiment as part of the Separation Management project. This was the third in a series of HITL experiments that examines new system Separation Management concepts and prototypes. ANG-E25 conducted the first experiment to investigate the effect of variable separation requirements within a sector and among aircraft, as well as a variety of automation tools to aid controllers in managing variable separation requirements (Sollenberger, Willems, DiRico, Hale, & Deshmukh, 2010). The second experiment evaluated issues that included (a) the location and format of the Conflict Probe (CP) notifications on the controller workstation, (b) the replacement of the current Radar (R-Side) and/or Radar Associate (RA-Side[1]) controller workstation display with a 30-inch commercial-off-the-shelf (COTS) monitor, and (c) pointing device alternatives to the existing controller workstation trackball (Higgins, Willems, Johnson, & Zingale, 2012; Zingale, Willems, Schulz, & Higgins, 2012). ATO En Route Requirements (AJV-731) sponsored the HITL experiment reported in this document and other activities to guide decisions on future En Route Automation Modernization (ERAM) system modifications.

1.2 Purpose

In this HITL experiment, we evaluated the following issues: (a) the degree of CP Alert reliability required for operational acceptance of CP Alert display on the R-Side; (b) the effect on controller performance, attitudes, and behavior of algorithmic enhancements to the CP; and (c) the effect on controller performance, attitudes, and behavior of radar surveillance noise. Our results will assist AJV-731 to validate, modify, or further develop CP performance requirements for the ERAM system.

1.3 Experimental Design Overview

1.3.1 Design A - Required Alert Reliability for R-Side CP

En Route Air Traffic Control (ATC) workstations include an R-Side display to present a surveillance display of aircraft in the National Airspace System (NAS) and an RA-Side display to present data about the aircraft. In the operational configuration of CP, only the RA-Side receives notifications. CP information is therefore less accessible to R-Side controllers, particularly when they are working alone. Presenting CP Alerts directly on the R-Side display could aid in more strategic conflict resolution. However, it is also possible that the distraction of false alerts would be overwhelming and lead to an overall drop in controller performance.

The Separation Management Team has conducted several engineering studies (cf. Crowell, Fabian, Young, Musialek, & Paglione, 2011, 2012; Crowell & Schnitzer, 2013; Young, Schnitzer, DiDonato, & Yao, 2014) that showed that algorithmic enhancements and tightening of adherence bounds significantly improve CP Alert reliability. The algorithmic enhancements included a modification of the geometry of the adherence bounds (i.e., closer to the current position of the aircraft, the adherence bounds will be smaller and further into the future the adherence bounds become larger). As the system tries to predict aircraft position further into the future, there is an

[1] Controllers often refer to the RA position as the Data or D-side position.

increasing level of position uncertainty. The enhancements also included a change that forced the trajectory to start at the last known track position rather than at the flight plan position.

Another enhancement to the algorithm forces CP to recalculate its trajectories before releasing a CP Alert to the controller. In the legacy CP, the system calculates the likelihood of a conflict only once and then waits to release a CP Alert to the controller until it reaches the time to loss of separation set by the likelihood function. In the enhanced system, CP will recalculate the trajectories before releasing the alert and—if the likelihood is lower than that specified by the likelihood function for the current time to loss of separation—will delay releasing the alert. Despite the promising results of the engineering studies, we have not conducted an operational evaluation of the impact of the improvements on controller acceptance of CP integration into the R-Side working position. Requirements do not state which CP performance level will make CP integration on the R-Side position operationally acceptable.

Accounts from previous HITL experiments, demos, and discussions with operations representatives report conflicting controller opinions regarding the utility of displaying CP on the R-Side. Previous HITL experiments under the Separation Management program have not shown any substantial gains or losses in performance with R-Side and RA-Side CP relative to RA-Side Only. To safely implement CP on the R-Side, it is important to empirically establish the degree of alert reliability required to maintain or improve current performance levels and to be operationally acceptable. In the first component of the experiment (Design A), we planned to systematically vary the CP algorithm to determine the level of alert reliability required for (a) controller performance improvement with R-Side display, relative to RA-Side Only display, and (b) controller acceptance of R-Side CP display.

1.3.2 Design B - Algorithmic Enhancements

To test the effects of modeling improvements planned for near-term ERAM implementation, we planned to evaluate likelihood settings and algorithmic changes shown to improve trajectory accuracy (Crowell et al., 2011, 2012; Fincannon, Willems, & Masalonis, 2015). We also varied the presence or absence of radar surveillance noise. Absence of noise emulated a theoretical future air traffic setting where aircraft location data surveillance has become much more accurate, for example, through the implementation of Automatic Dependent Surveillance–Broadcast (ADS-B).

2. METHOD

This experiment examined the following issues: (a) the degree of CP Alert reliability required for operational acceptance of CP Alert display on the R-Side, (b) the impact of algorithmic enhancements to the CP, and (c) the impact of surveillance noise. The results of the experiment will assist the ATO En Route Requirements organization (AJV-731) to validate, modify, or further develop CP performance requirements for the ERAM system.

2.1 Participants

We recruited eight Certified Professional Controllers (CPCs) from the ERAM National User Team. We excluded controllers from the Air Route Traffic Control Center (ARTCC) that we used in the experiment (Indianapolis Center, ZID) to ensure an equal level of airspace familiarity across participants. Each participating controller was fully qualified at his or her facility and had a current medical certificate at the time of his or her participation. Participants spent four working days at the Research Development and Human Factors Laboratory (RDHFL) at the Federal Aviation Administration (FAA) William J. Hughes Technical Center (WJHTC) for this study.

Of the eight controllers, seven were male, and one was female. Their age averaged 50.13 years (*SD* = 3.94) and ranged from 42 to 54. Including military and FAA experience, each participant had spent an average of 26.04 years (*SD* = 2.88) working as an air traffic controller, with experience ranging from 21.83 years to 30.5 years. In the past 12 months, seven participants had controlled traffic all 12 months, and one participant had controlled traffic for 4 months (*M* = 11.00, *SD* = 2.83).

2.2 Research Personnel

Four engineering research psychologists conducted the experiment. They were responsible for briefing the participants, preparing and operating the simulators and data collection equipment, administering questionnaires, and leading debriefing discussions. One research psychologist and one Subject Matter Expert (SME), a retired controller, trained participants on the airspace, Standard Operating Procedures (SOPs), Letters of Agreement (LOAs), automation support tools, and all unfamiliar elements of the ERAM Computer-Human Interface (CHI). A team of research psychologists and SMEs developed the scenarios. Hardware and software engineers prepared equipment. Twelve simulation pilots operated pilot workstations, communicated with controller participants using standard ATC phraseology, and maneuvered simulated aircraft according to controller instructions.

2.3 Materials

At specific times during the study, we asked controllers and observers to complete paperwork. The times for this were (a) during the initial intake prior to completing any experimental runs, (b) following each experimental run, and (c) at the end of the study after completing all experimental runs. The following subsections address each piece of paperwork.

2.3.1 Informed Consent Statement

Each participant read and signed an informed consent statement (Appendix A) before participating. The statement described the general purpose and procedures for the study and informed participants of their rights and responsibilities as volunteers.

2.3.2 Biographical Questionnaire

Each participant completed the biographical questionnaire (Appendix B) before participating. We used this questionnaire to collect basic demographic information, including age, gender, and ATC experience.

2.3.3 Post-Scenario Questionnaire

After each scenario, participants completed a Post-Scenario Questionnaire (PSQ; Appendix C). The General section of the PSQ contained questions about participants' perceived level of performance, workload, situation awareness, and scenario difficulty. The CP section of the PSQ contained questions about participants' perceptions of CP performance during the scenario. The questions solicited estimates of how accurate, timely, useful, and distracting participants considered CP information. The PSQ contained open-ended questions, where participants could make additional comments relevant to the scenario.

2.3.4 Exit Questionnaire

After participants completed all of the scenarios, they responded to an Exit Questionnaire (Appendix D). The first section of this questionnaire contained items relating to the effectiveness of the training and the realism of the traffic scenarios, ATC equipment, pilot interaction, and the overall simulation. The second section of the questionnaire contained items relating to the CP. Questions in this section included estimates of CP performance (mirroring the items in the PSQ); ratings of CP Alert reliability, usefulness, and distraction in different circumstances in the field; and ratings for individual conditions and interface features. As with the PSQ, each section of the Exit Questionnaire included open-ended questions, which allowed participants to make additional comments about the experiment.

2.3.5 Over-the-Shoulder Rating Forms

We assigned a dedicated SME with ATC experience to each participant to observe the participant during each run. The SME observer rated the participant on various dimensions of performance using an Over-The-Shoulder Rating Form adapted from instruments the Human Factors Branch has used in past experiments (Sollenberger, Stein, & Gromelski, 1997). To eliminate confounds between inter-rater differences and the experimental conditions, the same observer rated a participant throughout all runs. See Appendix E for the Over-The-Shoulder Rating Form.

2.4 Facilities & Equipment

We conducted the experiment at the RDHFL. This facility supports high-fidelity ATC simulations and was the site for the two previous HITL experiments in the Separation Management program. Equipment for two sectors – each consisting of two controller workstations (one for an R-Side position and one for a RA-Side position) and associated audio/video recording equipment—was available in each of two experiment rooms, for a total of four sectors. The simulation pilot workstations were in a separate room within the RDHFL.

2.4.1 Simulation Environment

We created a high-fidelity ATC simulated environment using the Distributed Environment for Simulation, Rapid Engineering, and Experimentation (DESIREE); the Simulation Driver and Radar Recorder (SDRR); the Target Generation Facility (TGF); and the ERAM Evaluation System (EES). The WJHTC also developed TGF to model the performance of individual aircraft. Using pre-determined flight plans, simulation pilot inputs, and aircraft performance characteristics, TGF generated realistic radar tracks and other aircraft data to display through DESIREE. ERAM contractors developed EES to simulate ERAM functionality with maximum fidelity. Its logic was identical to the operational version of ERAM, but its architecture allowed researchers to turn off modules and change parameters. Collectively, this software provided a highly realistic environment for participants and robust data collection capability for researchers.

2.4.1.1 Hardware

Figure 1 shows the simulation environment in which participating controllers worked during the experiment. The following sections will discuss the controller workstations, the Workload Assessment Keypad (WAK), the communication system, the pilot workstations, and the functional Near-Infrared Spectroscopy (fNIRS) equipment.

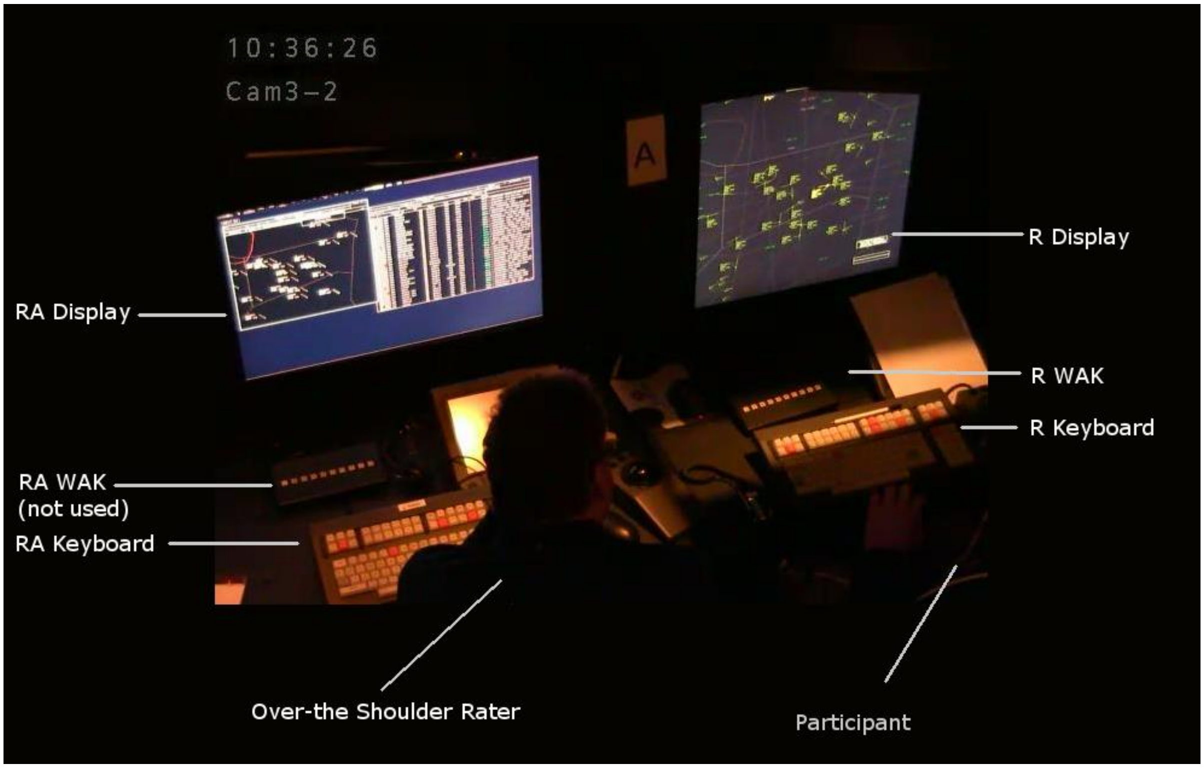

Figure 1. Simulation environment used in the experiment.

2.4.1.1.1 Controller Workstations

At each sector, we positioned R-Side and RA-Side controller workstation displays side by side. The R-Side workstations had a radar display, keyboard, Keypad Selection Device (KSD), and trackball. The RA-Side workstations had a display containing aircraft data—the Aircraft List (ACL) View—and Graphic Plan Display (GPD) depicting trajectory-based data, a keyboard, and a trackball. The keyboard, KSD, and trackball are custom devices manufactured by Cortron. The R-Side monitor was a BARCO 29-inch LCD with a resolution of 2048 x 2048 pixels. The D-side monitor was an EIZO 30-inch LCD with a resolution of 2560 x 1600 pixels.

2.4.1.1.2 Workload Assessment Keypad

We measured controllers' subjective workload in real time using a Workload Assessment Keypad (Stein, 1985). The WAK is a small device with 10 buttons (labeled 1 to 10). It connects through a computer to DESIREE that controls device timing and records data. At preset times defined by the researcher, the buttons illuminated, and the device beeped to prompt participants for a workload rating response. A rating of 1 indicated *very low workload,* and a rating of 10 indicated *very high workload* (refer to anchors in Appendix F). The buttons remained illuminated until the participant responded or until a researcher-defined interval had elapsed. If participants did not respond within the interval, the system recorded a code to indicate a non-response.

2.4.1.1.3 Communication System

We used a simulated Voice Switching and Control System (VSCS), similar to what is currently in the field. The VSCS enabled Push-To-Talk (PTT) voice communication between controllers and simulation pilots. The hardware included individual switchboxes, headsets, microphones, and PTT handsets. The equipment recorded the time and duration of all PTT transmissions. The microphones were always hot, which allowed researchers to record ambient communications.

2.4.1.1.4 Simulation Pilot Workstations

There were 12 simulation pilot workstations. Each workstation had a computer, monitor, keyboard, mouse, and communications equipment. The software that the simulation pilots used included a situation display of the airspace and traffic, a list of aircraft assigned to the pilot with all relevant flight data, and an interface to input flight plan changes.

2.4.1.1.5 Functional Near-Infrared Spectroscopy Equipment

We used a functional Near-Infrared Spectroscopy (fNIRS) Model 1100 system (fNIR Devices, Potomac, MD) to record prefrontal cortical activity and assess cognitive workload. This system assessed brain activity by using wavelengths of light to measure ratios of oxygenated and deoxygenated hemoglobin. The equipment includes a headband (see Figure 2) that contains a sensor package. We connected the headband to the participant's head with straps, and we connected the sensor package in the headband to a computer with two clips (see Figure 3). The sensor package in the headband included a wired forehead sensor pad that contained four light sources and 10 detectors to measure activity from 16 locations. We used alcohol swabs to sanitize the headband after each use.

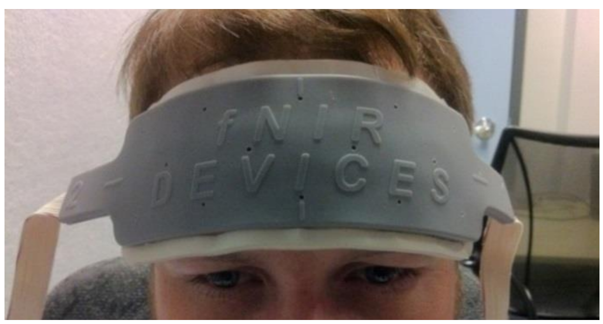

Figure 2. The fNIRS headband.

Figure 3. Clips to attach sensors in the fNIRS headband to the computer.

2.4.1.2 Software

2.4.1.2.1 DESIREE

WJHTC software engineers developed DESIREE to emulate en route and terminal workstation functionality. It provided a platform for flexibly modifying the system interface and capabilities to allow researchers to test novel concepts and procedures.

2.4.1.2.1.1 Experiment Control

The DESIREE Supervisor interface (see Figure 4) allowed the experimenters to set the conditions (values for independent variables), start and stop the simulation, and monitor the health of the systems. DESIREE used these settings to automatically generate run labels for any of the recording files it generated.

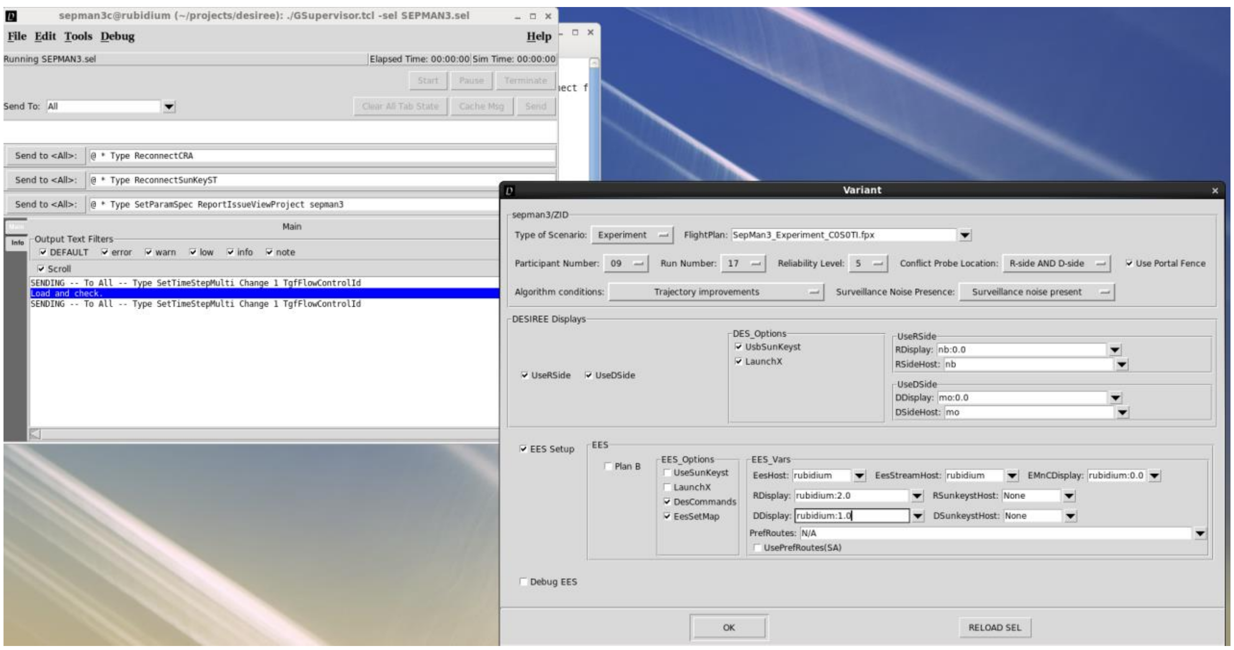

Figure 4. The DESIREE Supervisor interface showing the dialogue box used to set the values for independent variables, participant and run numbers, and simulation environment variables.

During the simulations, the experimenters monitored the DESIREE Supervisor interface (see Figure 5) to determine if issues occurred that required assistance from the development team. The DESIREE Supervisor interface provides the experimenters a quick view of the health of each of the computers involved in the experiment, warnings or errors generated by systems or subsystems, and the configuration settings.

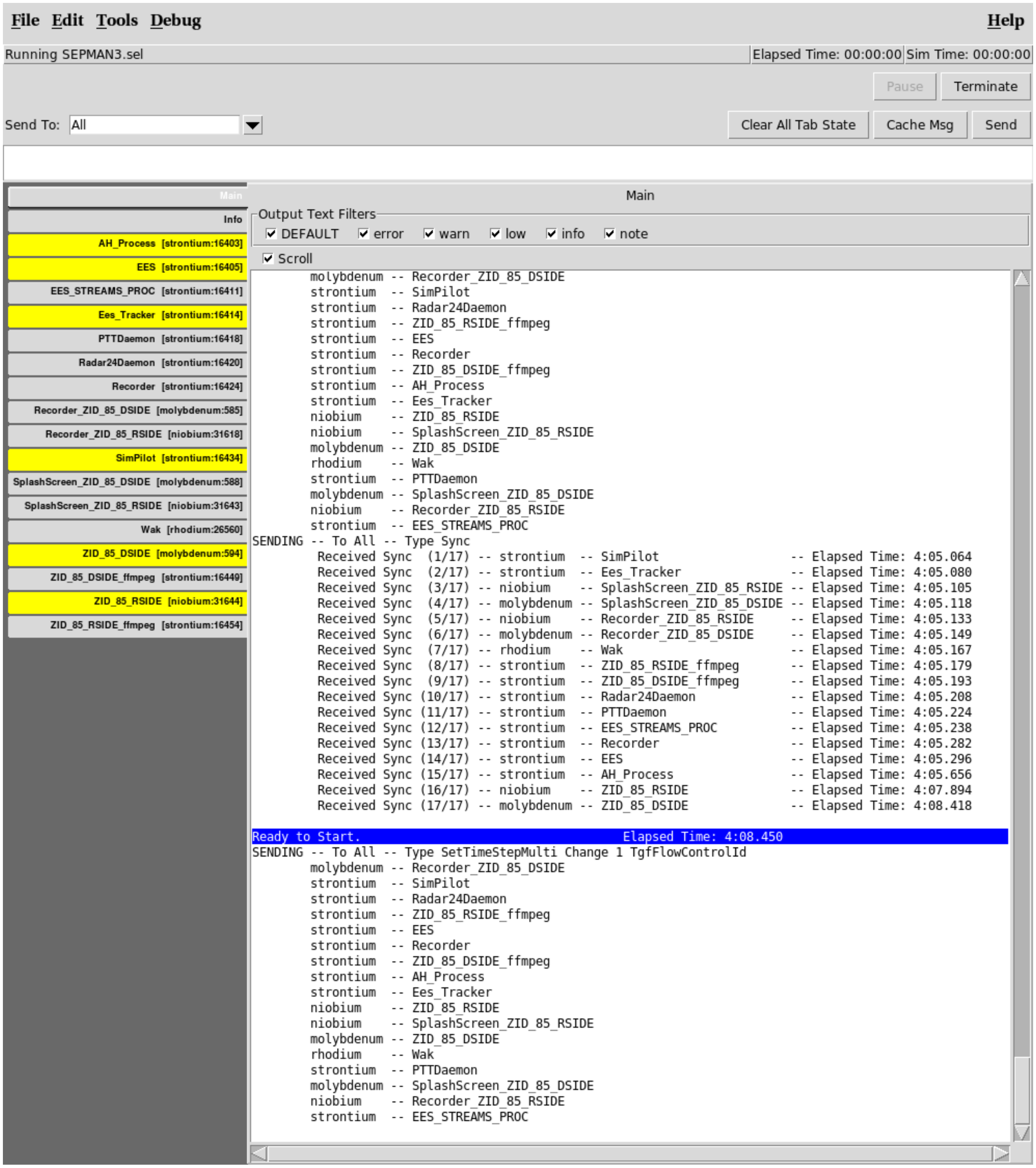

Figure 5. DESIREE Supervisor main tab provides experimenters with the ability to start and stop the simulation and indicates if warnings (yellow) or errors (red) are present in any of the processes.

2.4.1.2.1.2 ERAM CHI

The simulated controller interface emulated a possible future release of ERAM. Most features of the ERAM Radar and Radar Associate computer human interface were available (FAA, 2015a, 2015b). Where necessary, we provided new features to accommodate the integration of CP on the Radar display. The location of the CP varied in accordance with the experimental conditions. When the CP appeared on the R-Side, it included red alerts and airspace alerts (see Figure 6). In the Legacy/R-Side condition, the CP presented yellow and muted alerts (the B-R1 condition; see Table 3, Section 2.8.2). Yellow alerts activate when CP predicts the conformance boxes surrounding the aircraft trajectories to come within minimum separation values. Red alerts occur only when CP predicts the trajectory centerlines to violate separation. A muted alert occurs when the alert condition—red or yellow—will occur only if an aircraft proceeds on a "not yet cleared" portion of its trajectory, such as an assigned altitude. In all conditions, all alerts appeared on the RA-Side in the ACL and on the GPD. In spite of this intent, there were technical complications that influenced the display of alerts on the R-Side that we describe in Section 2.10.

Controllers had the capability to acknowledge alerts, show all alerts, or show specific alert pairs. When an alert occurred, the alert portal appeared attached to the Full Data Block (FDB) for each affected aircraft. The FDB included a red background and a white numeric indicator, denoting the total number of alerts for the aircraft. When controllers acknowledged the alert, the portal background turned transparent. Each predicted conflict had an entry in the Conflict Probe Alert View (CPAV; see Figure 7).

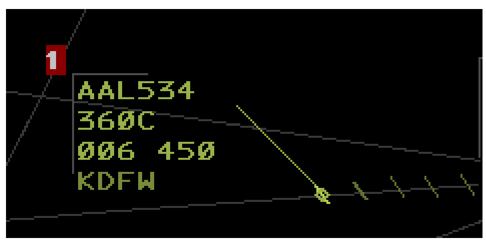

Figure 6. ERAM aircraft representation, including a Full Data Block, displaying the portal fence (gray line on the left and top of the Full Data Block) and the Safety Portal for a single potential conflict.

Figure 7. Conflict Probe Alert View showing two aircraft-to-aircraft alerts (in this case, the predicted loss of separation will occur in 5 minutes between ASQ3960 and DAL360) and one aircraft-to-airspace alert (in this case, JIA136 is under control of Sector 83 and loss of separation between the aircraft and a Special Activity Airspace [SAA] will occur in 1 minute).

When implemented, controllers will be able to set a look-ahead time via a Probe Alert Toolbar, which included buttons distinguishing aircraft-to-aircraft alerts from aircraft-to-airspace alerts. For this experiment, we fixed the look-ahead time at 10 minutes for aircraft-aircraft conflicts and 5 minutes for aircraft-airspace conflicts.

When controllers wanted to display the trajectories of a potential conflict, they either left-clicked on the safety indicator in the FDB or CPAV or used a center-click to show all potential conflicts for an aircraft (see Figure 8).

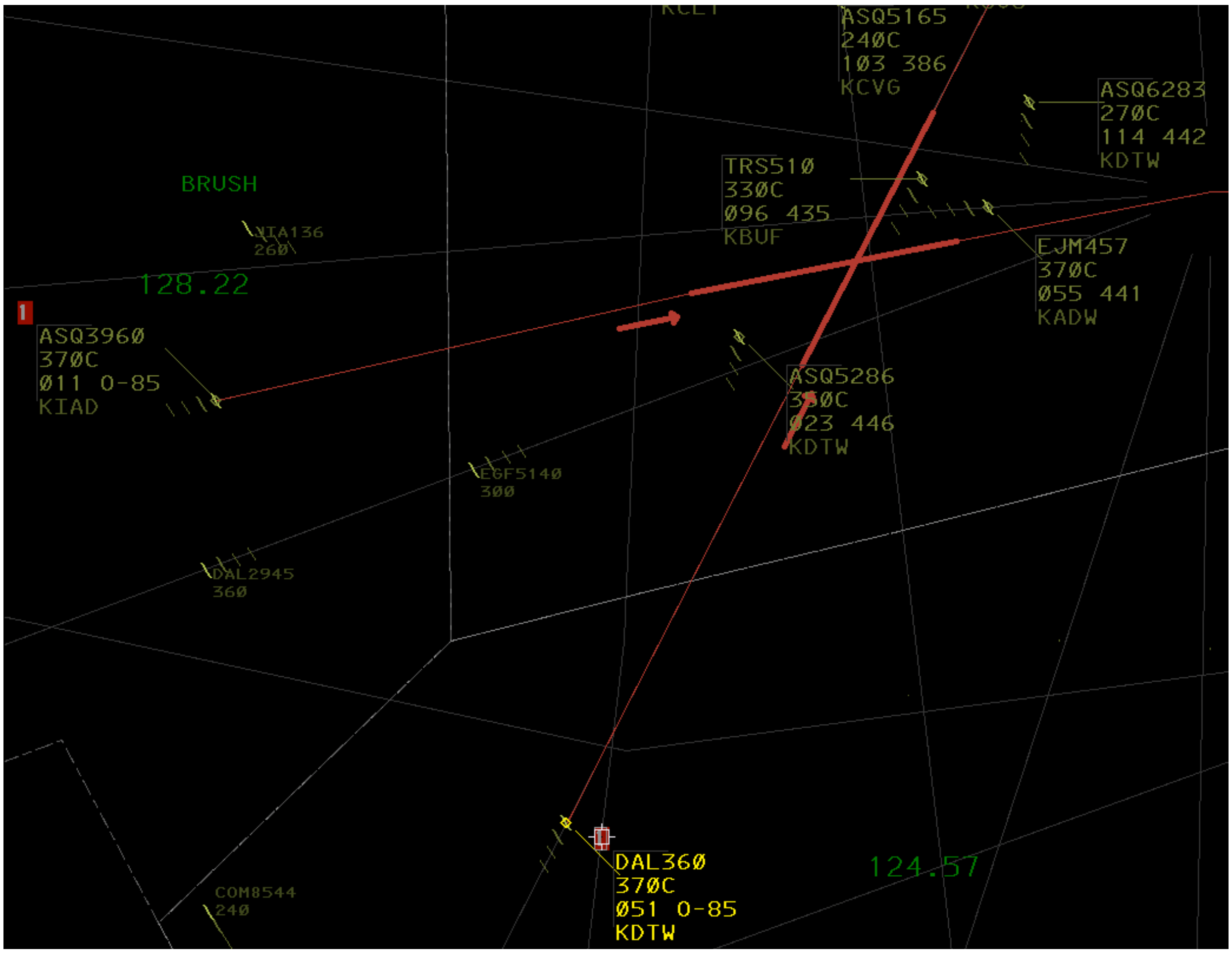

Figure 8. Trajectory display of a potential conflict between ASQ3960 and DAL360. In this case, the controller clicked on the safety indicator in the FDB of DAL360. The display shows the trajectories in red: The arrows indicate the direction of flight, and the trajectory segments indicate where the predicted loss of separation will occur in bold red lines.

2.4.1.2.2 TGF

The TGF software ingested airspace, route, aircraft characteristics, voice communications settings, and flight plan data. The TGF software generated aircraft position data and provided a software interface to start and stop the simulation, exchanged data with SDRR and DESIREE, and allowed simulation pilots to monitor airspace and maneuver aircraft. The software accepted simulation pilot commands that DESIREE generated to manipulate aircraft in sectors and facilities other than studied during the experiment.

2.4.1.2.3 SDRR

The SDRR software received aircraft position updates from TGF and generated surveillance data that included surveillance noise when necessary. The software fed the surveillance data to EES that accepted the data as if it came from true surveillance sources. The SDRR software also emulated facilities adjacent to the ZID ARTCC. In this simulation, those facilities consisted of Airport Traffic Control Towers (ATCTs), Terminal Radar Approach Control Facilities (TRACONs), and ARTCC HOSTs. To inject controller commands and inter-facility messages, the SDRR software also accepted DESIREE generated messages that it passed on to the ERAM software in EES. The SDRR software provided the experimenters an interface to monitor the transfer of messages between the emulated systems and EES (see Figure 9).

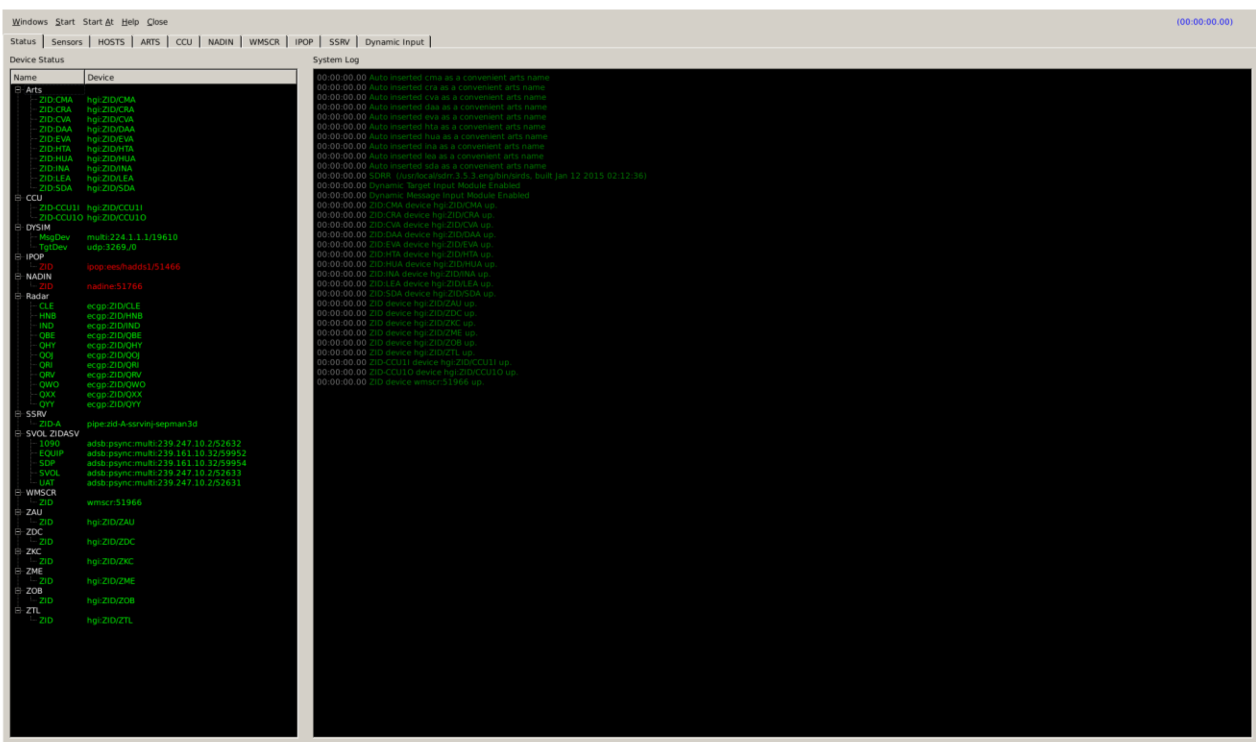

Figure 9. SDRR status view showing the status of simulated Automated Radar Tracking Systems (ARTS), ARTCC, and radar facilities. The connections to external automation (IPOP) and National Airspace Data Interchange Network (NADIN) do not become active until the simulation starts.

2.5 Airspace

Previous experiments (cf. Zingale et al., 2012) had not found statistically significant changes with CP on the R-Side, as compared to CP on the RA-Side Only. This may be because the presence of CP on the R-Side display truly makes no difference. It is also possible that the negative and positive consequences of R-Side display are only significant in high-workload circumstances, which is when reliable information about conflicts is of most value and the distraction of false alerts is most costly. Assuming that the latter was the case, it was important to test the effect of the different CP algorithms and to display alternatives in an environment with high traffic and a high alert rate.

We used data from the MITRE Corporation's CAASD Analysis Platform for En Route (CAPER) to identify potential sectors (cf. Kell & Chen, 2004). The CAPER CP differs from ERAM's CP algorithm but is qualitatively similar. The database of CAPER probe events was also more comprehensive than similar databases for the User Request Evaluation Tool (URET) or ERAM probe. MITRE/CAASD provided the average number of CAPER probe-predicted "separation events" for every sector in the contiguous United States between 1400 GMT and 2400 GMT every day for the first six months of 2011. They defined a loss of separation event as a pair of aircraft predicted to come within seven miles horizontally and 1,000 feet (at or below flight level 410) or 2,000 feet (above 410) vertically. These events did not perfectly correspond to ERAM CP Alerts but should have reflected sector-by-sector differences in the rate of potential conflicts and associated alerts.

From the 878 sectors in the 20 ARTCCs in the contiguous United States, we selected ZID85 from Indianapolis Center as the airspace for this experiment, since it had one of the highest rates in the NAS of separation events. Other sectors with a higher rate of loss of separation events were less ideal for use in an experiment focusing on aircraft conflicts. Some of the higher ranking sectors were narrow and long and, therefore, had less potential for a variety of conflict angles. ZID85 also is often combined with the sector or sectors directly above it with the same lateral boundaries, which gave us more flexibility in scenario design. We combined ZID85, a high-altitude sector, with ZID79 and ZID95, the intermediate and super high sectors, respectively, overlying it at the same lateral limits. Throughout each run, the Buckeye Special Activity Airspace (SAA) was active; it overlaps the northwest corner of ZID85. We depicted its location on the R-Side display, and CP presented aircraft-to-airspace alerts for flights whose trajectories intersected the SAA. Figure 10 provides an illustration of this sector, the SAA, and the key fixes and airports in and near the sector.

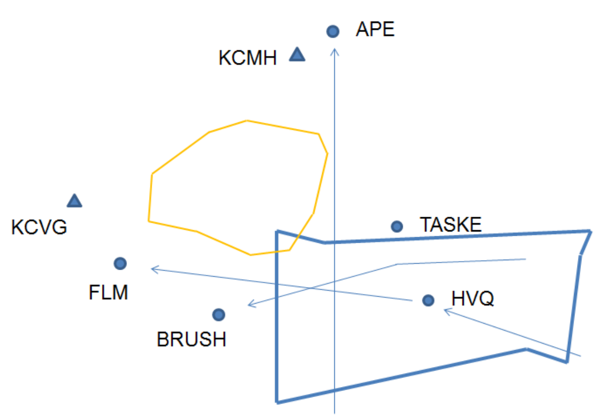

Figure 10. Sector ZID85 with key fixes and flows and the SAA.

2.6 Traffic Scenarios

We based all traffic scenarios in this experiment on real traffic samples from ZID. We created two unique training scenarios, each of which ran for 30 minutes. The first began at 33% of the sectors' (i.e., ZID85, ZID79, and ZID95) combined Monitor Alert Parameter (MAP), which we treated as 19 aircraft for the combination of these three sectors because it is the highest of the individual MAPs of the sectors. Once the traffic level increased to 50% of the MAP value, it remained for the duration of the run. We used this scenario to familiarize the controllers with the airspace, procedures, and ERAM interface features that are common across designs.

We created one 60-minute scenario as the baseline for the practice scenarios. From this baseline, we altered the start times by cutting several minutes from the beginning of selected versions of the 60-minute scenario so that each resulting scenario would present a slightly different window within the master 60-minute period, with a core period identical across all experimental scenarios to maximize comparability between runs. We also scrambled aircraft call signs and changed some destinations from run to run. Four practice scenarios resulted, and we used three of the four during the practice runs. The purpose of the cutting and scrambling process was to minimize participants' opportunities to overlearn the scenarios, while maintaining comparability across scenarios for subsequent data analysis. Our practice scenarios began at 33% of MAP, increased over the first 5-7 minutes to 125% of MAP, and decreased to approximately 75% of MAP.

With the same approach of using a single baseline scenario and altering start times, call signs, and destinations, we created 16 experimental scenarios, 12 of which we used during the experimental runs. The baseline scenario began at 33% of MAP and ramped up over the next 5-10 minutes to 125% of MAP, later decreasing gradually to 66% of MAP and remaining until the end of the scenario.

2.7 Conflict Probe Alert Reliability

We used an adjustment to the conformance bounds to manipulate CP Alert reliability. As required by Design A (see below), the CP Alert reliability of the A-R2 and A-RA2 conditions needed to match CP Alert reliability under traditional settings. Because the CP in Design A uses algorithmic enhancements in a simulation without radar noise, 2.5 nautical miles (NM) lateral and 1.5 NM longitudinal conformance bound with a 10-20 likelihood setting (explained in Section 2.8.2) should produce a more accurate CP in Design A than what one would traditionally observe with these settings. Previous work examined this issue and found that a 2.5 NM lateral and 2.0 NM longitudinal conformance bound with a 4-8-20 likelihood setting and algorithmic enhancements would produce a level of CP Alert reliability that was equivalent to legacy conditions (Fincannon et al., 2015). Table 1 provides a list of the conformance bound settings and CP Alert reliability that correspond to each CP Alert reliability level with red and yellow alerts. It is important to note that we use a human factors formula to define CP Alert reliability, and Formula 1 illustrates this equation. We had a problem with the simulation configurations that caused us to run all conditions as a 2.5 NM lateral and 2.5 NM longitudinal conformance bound with a 10-20 likelihood setting. We should note that there were technical complications influencing the display of alerts on the R-Side that we describe in Section 2.10.

Table 1. Design A - Relationship between Conformance Bound Settings and CP Alert Reliability with Algorithmic Enhancements, No Surveillance Noise, and a 4-8-20 Likelihood

Conformance Bounds and CP Alert Reliability	CP Alert Reliability Level				
	1	2	3	4	5
Lateral Conformance Bound	2.5	2.5	1.0	1.0	0.1
Longitudinal Conformance Bound	2.5	2.0	2.5	1.25	0.1
CP Alert Reliability %	59.9%	61.8%	67.3%	74.6%	85.7%

$$reliabilty = \frac{H + CR}{H + CR + M + FA} \tag{1}$$

where H = hits, CR = correct rejections, M = misses, and FA = false alarms.

2.8 Experimental Design

2.8.1 Design A - Required Alert Reliability for R-Side CP

We used Design A to examine the interaction between CP Alert reliability and display location. The original experimental design called for systematically varying CP settings to determine the level of CP Alert reliability required for R-Side implementation. Because of simulation configuration errors, we did not manipulate CP Alert reliability as designed. More detail on the actual conditions run appears in Section 2.10.

We developed five sets of lateral and longitudinal conformance bounds to create five unique configurations. These configurations had an approximately equal rate of valid alerts, but they can increase or decrease the number of false alert rates. Based on previous research by the Modeling and Simulation Branch (ANG-C55), we conducted a simulation with large-scale scenarios and analyzed CP data to develop an understanding of how conformance bound settings influenced the alert reliability of the CP (Fincannon et al., 2015). We selected one of the CP Alert reliability levels (level 2) to match the CP Alert reliability (see Formula 1) of the legacy algorithms reported by Crowell et al. (2012). Another (level 1) was lower than this reference level, and three (levels 3, 4, and 5) were higher than the current level. Based on preliminary analysis by Fincannon et al. (2015), the change from level 1 to 5 consisted largely of reduction in false alerts rather than of missed alerts.

Controllers worked alone with an R-Side and RA-Side display. We designed eight testing conditions varying CP Location and CP Alert reliability level, as shown in Table 2. In the three RA-Side conditions, we displayed CP only on the RA-Side at the reference reliability level (A-RA2), a higher reliability level (A-RA3), and the highest reliability level (A-RA5). In the five R-Side conditions (A-R1 to A-R5), we displayed CP on the R-Side and RA-Side at all five CP Alert reliability levels. Each participant conducted one test run for each of the eight conditions. We randomly assigned the run order across conditions. To allow the testing of higher reliability levels, all conditions in Design A had the algorithmic improvements (discussed in Sections 1.3.1 and 2.7) and did not have surveillance noise.

Table 2. Design A - Evaluation of Required Accuracy for R-Side CP with Improved Algorithm and No Surveillance Noise

	CP Alert Reliability Level				
CP Location	**1**	**2**	**3**	**4**	**5**
RA		A-RA2	A-RA3		A-RA5
R + RA	A-R1	A-R2	A-R3	A-R4	A-R5

We planned a one-way comparison to examine the five levels of the R+RA (A-R1 to A-R5) presentation of the CP. We also planned to examine the effect of CP Reliability and CP Location in a 3 x 2 design: three levels of Reliability (2, 3, and 5) and two levels of Location (R + RA vs. RA Only).

2.8.2 Design B - Algorithmic Improvements

Design B evaluated the impact of Algorithmic Improvements, CP Location, and Surveillance Noise. As in Design A, controllers worked alone with an R-Side and RA-Side display. The Legacy CP condition was designed to use the currently fielded CP configuration. This configuration includes a 10-20 likelihood function. The likelihood function (Crowell & Schnitzer, 2013) controls when alerts are displayed based on the computed likelihood that the aircraft pair will lose separation if nothing is done. For a 10-20 function, the CP will provide an alert on a situation with likelihood of separation loss of 1 (complete certainty) at 20 minutes before the predicted loss and will report a situation with a likelihood of 0 at 10 minutes prior to the time it is predicted to happen.[2] The algorithm determines the timing of alert display for likelihoods between 0 and 1 by a linear interpolation between these two points. For example, the system will display an event with a 0.5 likelihood of separation loss—halfway between 0 and 1—at 15 minutes beforehand, halfway between 10 and 20 minutes before predicted loss of separation.

The configuration we intended to use for Legacy conditions corresponds to legacy conformance bound settings of 2.5 NM (lateral) and 1.5 NM (longitudinal). In the Enhancements (Improved Algorithms) condition (see Young et al., 2014), we intended to use the likelihood function recommended by Crowell and Schnitzer (2013) of 4-8-20. This function reports situations with 0 likelihood of losing separation at 4 minutes before the event; the likelihood is 0.9 will be reported at 8 minutes look ahead, and one with a likelihood of 1.0 at 20 minutes. For likelihoods falling between 0 and 0.9, the system interpolates linearly between the 4- and 8-minute points, and if the likelihood is between 0.9 and 1, the system interpolates linearly between the 8- and 20-minute points.

[2] The alert timing for a zero-likelihood loss of separation is a hypothetical point used to create the function. If CP does not expect a loss of separation, it does not provide an alert. With a 10-20 likelihood function, CP will alert an event with just greater than 0 likelihood at just above 10 minutes.

We manipulated the two levels of the algorithmic changes (see Table 3) across two presentations of the CP: the RA-Side Only and the RA-Side plus the R-Side. This manipulation allowed for a 2 (Legacy, Enhancements) x 2 (RA and RA + R) analysis of Algorithmic Improvements and CP Location. For all four runs compared in this analysis, Surveillance Noise was present.

Table 3. Design B - Evaluation of the Impact of Algorithmic Improvements, Surveillance Noise, and Position with Current Adherence Bounds

		Algorithmic Improvements	
Surveillance Noise	**CP Location**	**Legacy**	**Enhancements**
Yes	**RA**	B-RA1	B-RA2
	R + RA	B-R1	B-R2
No	**RA**		A-RA2
	R + RA		A-R2

To model surveillance noise, we replicated procedures by Thompson and Flavin (2006) for simulating radar error and used values from the Monte Carlo simulation to introduce jitter. We set radar jitter to 25 ft (0.004 mi), and we set azimuth jitter to 0.23º.

To examine the impact of Surveillance Noise, CP display location, and algorithmic improvements, we incorporated two conditions from Design A. We designed these conditions to include the Enhancement (Improved Algorithms) settings but not Surveillance Noise. This manipulation allowed for a 2 (Surveillance Noise and No Surveillance Noise) x 2 (RA and RA + R) comparison of Surveillance Noise and CP Location.

2.8.3 Design A & B - fNIRS Analysis

We used fNIRS as an objective physiological assessment of controller workload. We ran four participants at once but had access to only two fNIRS headbands, which limited the number of conditions in which we could use this equipment to assess workload. To control how this impacted the analysis, we selected a subset of conditions to focus on two general questions. First, we wanted to assess the impact of CP Location. We selected the two location conditions with the highest CP Alert reliability level from Design A (A-RA5 and A-R5) to address this question while holding CP Alert reliability constant. Second, we wanted to look for non-linear effects of CP Alert reliability on our dependent measures. This analysis called for including A-R1, A-R3, A-R5, B-R1, and B-R2, which allowed us to assess the impact of CP Alert reliability as manipulated by both algorithmic enhancements and conformance bound settings.

2.8.4 Dependent Measures (Designs A & B)

We collected a variety of measures to assess the effects of CP Alert reliability and location. These measures fell into the following categories:

- **Aggregate safety and efficiency** – These measures included the number of altitude, heading, and speed commands issued; the time and distance flown in the sector; the timing of handoffs; the number and duration of air/ground communications; the number and duration of CP Alerts; the number of operational deviations of various types; and the number of losses of separation.
- **Deviations** – We trained participants on a set of SOPs that represented a modified subset of actual SOPs for ZID85. Altitude restrictions were in effect for aircraft landing at Columbus (CMH) and Cincinnati (CVG), which had to be level at FL240 when exiting ZID85; and landers at Canton-Akron (CAK), which had to be at or descending to FL270 or below, direction of flight appropriate. Also, when handing off all exiting flights to neighboring sectors, we instructed controllers to use the correct frequency of the appropriate sector and to complete the transmission to transfer communications prior to boundary crossing as in real operations. In all runs, we placed the numbers and frequencies of each neighboring sector on the R-Side displays by using the drawing function. Also, we placed reference sheets at each position, showing the same information on a map and detailing the altitude restrictions.
- **Aggregate aircraft data interaction** – This category of measures includes the number of vector line changes, the number and duration of flight plan readouts, the number and duration of route displays, and the number and duration of CP trajectory display instances.
- **Scripted conflicts measures** – These are measures calculated solely on the specific conflicts that we designed into the scenario and on the aircraft involved. It included many of the measures (described above under "aggregate aircraft data interaction") as well as the timing of clearances issued relative to when the conflict would have occurred.
- **Subjective ratings** – These included WAK ratings and questionnaire responses measuring (a) controller workload, performance, situation awareness, and trust in automation; (b) perceived CP performance and usefulness; and (c) usefulness of the display formats and tools.
- **Physiological measures** – We measured participants' physiological responses to experimental conditions by assessing frontal lobe oxygenation as measured by fNIRS.

2.9 Procedure

2.9.1 General Schedule of Events

Eight current En Route controllers participated in the study. This included two groups of four participants. Each group participated over the course of four days. We provided training for a full day on Day 1. On Day 2 through Day 4, they participated in the testing sessions.

2.9.2 Initial Briefing

We began each four-day rotation by briefing the controllers on the purpose of the experiment, the schedule of activities, and the means by which we would collect data (e.g., fNIRS, WAK). The controllers each read and signed the informed consent statement after which an investigator and a witness signed the consent statement as well. The participants then filled out the Background Questionnaire. A research psychologist described the background, purpose, and conduct of the experiment, and a SME described the airspace, SOPs, and LOAs for the experiment.

2.9.3 Training

The initial training run introduced the controllers to the simulator equipment and interface. We instructed them on the novel ERAM features and simulation environment. They then completed a second training during which we introduced the features and procedures of individual designs and conditions and provided appropriate instructions during each run. Finally, there were three practice runs. During the training and practice runs, controllers rotated through workstations where they wore the fNIRS equipment a minimum of two times to familiarize themselves with it before the data collection runs. In all runs, the system prompted controllers for WAK ratings at 2-minute intervals.

2.9.4 Data Collection

There were 12 test sessions, one for each of the 12 designed conditions. We counterbalanced condition order by using a modified Latin Square Design. At the start of each session, we instructed each participant at which of the four positions to sit. Seating assignments were rotated primarily so that we could collect fNIRS data for the selected conditions and not need to move the fNIRS equipment. Therefore, when a participant was running one of the conditions selected for fNIRS recording, he or she sat at one of the two positions where the fNIRS equipment was available.

Participants provided WAK ratings at 2-minute intervals in the test runs. Immediately after each test run ended, they completed all of the relevant sections of the PSQ for the condition. After a break, the next test session began. Most testing days ended with a group discussion during which controllers could share their thoughts on the day's simulation runs. On the final day, the controllers completed the Exit Questionnaire. We held a final debriefing session in which the controllers shared their final thoughts about the simulation experience and the tools and procedures. We fully debriefed them about the nature of the experiment and answered questions.

A sample schedule for one team appears in Table 4. Designs A and B required a total of 12 test sessions, as depicted in Table 2 and Table 3. The schedule for Day 2 through Day 4 allowed for these 12 sessions and three makeup sessions.

Table 4. Schedule of Events

Day 1	Day 2	Day 3	Day 4
Welcome 8:00 - 8:15	Briefing 8:00 - 8:15	Briefing 8:00 - 8:15	Briefing 8:00 - 8:15
Airspace and Procedures Training 8:15 - 9:45	Test 8:15 - 9:15	Test 8:15 - 9:15	Test 8:15 - 9:15
Break 9:45 - 10:00	Break 9:15 - 9:30	Break 9:15 - 9:30	Break 9:15 - 9:30
Training Scenario 10:00 - 10:45	Test 9:30 - 10:30	Test 9:30 - 10:30	Test 9:30 - 10:30
Break 10:45 - 11:00	Break 10:30 - 10:45	Break 10:30 - 10:45	Break 10:30 - 10:45
Training Scenario 11:00 - 11:45	Test 10:45 - 11:45	Test 10:45 - 11:45	Test (makeup) 10:45 - 11:45
Lunch ***11:45 - 12:45***	***Lunch*** ***11:45 - 12:45***	***Lunch*** ***11:45 - 12:45***	***Lunch*** ***11:45 - 12:45***
Practice 12:45 - 1:45	Test 12:45 - 1:45	Test 12:45 - 1:45	Test (makeup) 12:45 - 1:45
Break 1:45 - 2:00	Break 1:45 - 2:00	Break 1:45 - 2:00	Break 1:45 - 2:00
Practice 2:00 - 3:00	Test 2:00 - 3:00	Test 2:00 - 3:00	Test (makeup) 2:00 - 3:00
Break 3:00 - 3:15	Break 3:00 - 3:15	Break 3:00 - 3:15	Caucus & Debrief 3:00 - 4:30
Practice 3:15 - 4:15	Caucus 3:15 - 4:30	Caucus 3:15 - 4:30	
Caucus 4:15 - 4:30			

2.10 Complications and Modified Experimental Design

During the data reduction and analysis phase of this experiment, we discovered that not all of the conditions ran as designed. As stated earlier, we manipulated CP Location and three types of variables designed to affect CP Alert reliability: parameter settings (conformance bounds and likelihood functions), surveillance noise, and algorithmic enhancements. The three types of algorithmic settings work by overriding the EES defaults. While developing the scenarios, settings, and simulation environment for this experiment, we verified that the settings achieved the desired results. However, during the later simulation development and shakedown phases, modifications to the platform had unintentionally affected the DESIREE-EES communications and prevented the overrides from working.

During initial phases of data analysis, the pattern of results did not correspond with expectations; that is, the improvements in reliability did not positively correlate to our initial findings. To investigate the discrepancy, we examined the simulation files to determine what settings were active during runtime. During the execution of the experiment, the researchers made the proper selections when launching each run (i.e., specified the correct CP settings for each run), but the DESIREE did not transfer the settings properly to EES. Therefore, the EES ran the default conformance bounds (2.5 NM lateral/2.5 NM longitudinal) and likelihood function (10-20) in all experimental conditions.

Appendix G shows the CP Alert reliability and display settings that we intended to run in each condition as well as the settings that were actually run. As a result of the differences between planned and actual settings, we could not conduct analyses by using the designed levels of CP Alert reliability. The conformance bound settings did not vary, and while we designed conditions with algorithmic enhancements to use a 4-8-20 likelihood setting, we always used a 10-20 likelihood setting. Outside of these problems, the algorithmic enhancements were correctly included or excluded as specified for each run, and the CP Location and Surveillance Noise manipulations did run as specified, so we were able to conduct analyses relying on differences in these variables. The CP Alert reliability differences between the two levels of algorithmic enhancements may have been smaller than planned (i.e., the modeling enhancements varied, but the likelihood function did not), but we deemed that participants might still be affected by the CP Alert reliability differences that did occur between these conditions. We analyzed this variable, but it might be harder to detect differences in our dependent measures between the conditions.

As a result of these complications, most of the analyses reported in Section 3 used only the 6 of 12 conditions that ran approximately as designed: A-R2, A-RA2, B-R1, B-R2, B-RA1, and B-RA2. Testing of overall Location effects for some dependent measures included all six conditions, comparing the three R-Side CP conditions against the three RA-only CP conditions. For testing Surveillance Noise effects, we were interested in the main effect of Noise and its interaction with Location. Therefore, we used the subset of four conditions where other variables, particularly algorithmic enhancements, were held constant: A-R2, A-RA2, B-R2, and B-RA2. To assess the effects of algorithmic enhancements, we were interested in the main effect of algorithmic enhancements and its interaction with Location. Therefore we used the four conditions where other variables, in particular Noise, were held constant. These conditions were B-R1, B-R2, B-RA1, and B-RA2.

In addition to the issues that caused some of our Independent Variables (IVs) not to be presented as we intended, technical simulator issues sometimes caused problems with the handling of individual flights. For example, sometimes an aircraft that was designed to traverse the experimental sector during a scenario did not appear on DESIREE. This type of occurrence was occasional, but the most prevalent problem related to handoffs. A common handoff issue was the auto-handoff function attempting to hand off aircraft entering or exiting ZID85 to an incorrect sector. Another handoff issue was that at times the participant was unable to use the automation to accept or issue a handoff for certain aircraft. Experimenters, observers, simulation pilots, and TGF personnel kept careful records of which flights were affected and in which runs. During data reduction, we augmented this list by querying data, such as controller and simulation pilot commands, to identify other flights that were subject to simulation issues. For example, a flight for which a participant or observer entered "/OK" was most likely not handing off properly to the experimental sector. For each analysis involving individual aircraft, such as time/distance in sector and operational deviation data, we excluded flights with anomalous behavior of a type we believed could affect the validity of the analysis.

A final aspect of the experiment that did not run as intended was the display of muted alerts. Our experimental design called for showing muted alerts on the R-Side Only in the Legacy algorithm condition, B-R1. In fact, this was the only condition in which the participants saw alerts in the muted red (and yellow) colors on the R-Side. However, due to a setting in DESIREE, alerts that should have been muted red did display on the R-Side in all R-Side CP conditions, but in the same color as regular red alerts. The impact of this error to our results is less of an issue compared to the issue involving conformance bounds and likelihood settings. Participants often realized early in their participation that some of the red CP Alerts displayed on the R-Side were clearly for flights at different altitudes. Participants seem to have been able to learn to ignore this factor aside from making a few comments about it. Questionnaire responses regarding trust in the Probe were not inordinately low. During the experiment and in the replays we viewed during data analysis, we saw little to no evidence of participants taking control actions as a result of alerts that were clearly false.

Following the discovery of these issues, the DESIREE team identified and implemented the software and configuration changes necessary for EES to read in and activate the selected parameter settings. We have conducted runs of validation scenarios to demonstrate that the system now implements the settings properly when using the latest version of the simulation platform and affect CP Alert reliability in the expected direction. If we will use this platform for future research that requires the manipulation of CP Alert reliability, the conditions will run as designed. See Section 4 for further coverage of the implications of these issues for the current experiment and for future research.

3. RESULTS

For each analysis reported in this section, except where noted, we analyzed the effects of Location, Noise, Algorithm, and/or key interactions between these variables. The specific analyses conducted depended on the nature of the data, such as whether we expected a manipulation to influence the dependent measures, and the availability of sufficient data to conduct the analysis.

We examined the effect of algorithmic enhancements (Legacy vs. Enhancements) and CP Location (RA-Side Only vs. R- and RA-Side). These analyses used the four conditions from Design B—the set of conditions in Table 5 without radar surveillance noise, thereby holding that variable constant.

Table 5. Subset of Conditions - CP Algorithm by CP Location Analysis

	Algorithmic Improvements	
CP Location	**Legacy**	**Enhancements**
RA	B-RA1	B-RA2
R + RA	B-R1	B-R2

We also examined the effect of Radar Surveillance Noise (Present vs. Absent) and CP Location (RA-Side Only vs. R- and RA-Side). This analysis includes the set of conditions in Table 6 that included algorithmic enhancements, thereby holding the Algorithm variable constant.

Table 6. Subset of Conditions - Radar Surveillance Noise by CP Location Analysis

	Radar Surveillance Noise	
CP Location	**Present**	**Absent**
RA	B-RA2	A-RA2
R + RA	B-R2	A-R2

For data collected continually during the runs, such as WAK ratings, fNIRS data, deviations, and controller and pilot commands, to make the data from the various experimental conditions as comparable as possible, we used only a "core" period from each run. Our method for creating non-identical but comparable scenarios (see Section 2.6) by "cutting" the longer baseline scenario to start the initial runs from different points in that scenario resulted in a 35-minute core period that fell between minutes 13 and 48 in the baseline scenario. We cut this period to 34 minutes for analyses and used 2-minute intervals to remove the interval in each run that was only halfway in the core period. The core period contained the set of flights and events occurring between minutes 4 and 48 in all runs regardless of the cut point from the baseline scenario at which the run began. We used this period because in the first two minutes of each scenario, as noted in Section 2.9.4, the screens were blank while we recorded fNIRS baseline data. For the second 2-minute interval, we deemed that participants were still developing their traffic picture and settling into the problem. At the end of each run, participants entered the final WAK rating at the 50-minute mark. We informed them that the scenario was over—which could bias their instantaneous workload ratings and other experiment data, as they were no longer "working" at the instant they made the last rating. Also, the two minutes immediately preceding the end of the run could be biased by an "end spurt"—or conversely by lower motivation to fully perform all tasks—due to the knowledge that the run was almost over. The period we used for the relevant analyses corresponded to minutes 4 through 39 of the scenarios starting at minute 9 of the master scenario, and so on up to minutes 13 through 48 of the scenarios starting at minute 0 of the master.

For the continually collected data—except where noted, we conducted multiple regression analyses using the instantaneous aircraft count in the sector and the number of minutes into the run (time on task), as covariates. We entered these variables, along with the independent variable(s) of interest for the analysis, into the regression equation as fixed effects. We treated the Participant variable as a random effect to create a within-subjects model. For some dependent variables, we used Unique Run variable nested within the Participant variable as the random effect to create a multilevel within-subjects model on the theory that variance unique to the run itself, such as seating assignment or time of day, should be controlled for. However, for some variables, a multilevel approach did not add reliability to the model and could inflate the possibility of Type II errors, thereby failing to detect a true effect. For all applicable regression analyses, we conducted a preliminary comparison of the log likelihood values for the multilevel and non-multilevel analyses according to the guidelines outlined by Tabachnick and Fidell (2007), and report the results of the more appropriate model.

In this report, we present figures or tables, and descriptions, only of effects that were statistically significant at the .05 level. We selected trends where $p < .10$ or where we deem the pattern of results particularly operationally meaningful. Error bars on bar charts represent the standard error of the mean. We show more detailed results in the Appendices.

3.1 Workload Assessment Keypad

For all WAK rating analyses, as with most of the analyses described in this section, we assessed the main effects of Location, Noise, and Algorithm, as well as key interactions between them. We conducted multiple regression analyses by using each individual WAK rating as a unit of observation. We used the instantaneous aircraft count in the sector, at the minute mark when the WAK prompt activated, as a covariate. We also used the number of minutes into the run at which the interval began (time-on-task) as a covariate. These variables, along with the independent variable(s) of interest for the analysis (Location, Location/Algorithm, and Location/Noise), were fixed effects in the regression equation, and the Unique Run variable nested within the Participant variable was a random effect. For WAK data, our preliminary comparison of this model (with a non-multilevel within-subjects model where *participant* was the only random effect) showed that the multilevel model was more reliable for each of the three analyses described in this paragraph.

We first analyzed the mean WAK ratings as a function of Location, using the six conditions for each participant where the parameters of the conditions run were approximately as planned. The one-way analysis of Location (the data from the three R- and RA-Side conditions vs. the RA-Side Only conditions) did not show statistical significance. The analysis of Location and Algorithm did not reveal any significant main effects either, and the interaction was not significant. Likewise, the analysis of Location and Noise did not reveal any significant main effects, and the interaction was not significant. See Appendix H for detailed WAK mean rating results.

We assessed the effect of task load on subjective workload by computing the correlation between the instantaneous count of aircraft present in the sector at the minute each rating was taken, and the numerical value of the rating. Figure 11 is a scatterplot of each WAK rating and its associated level of traffic.

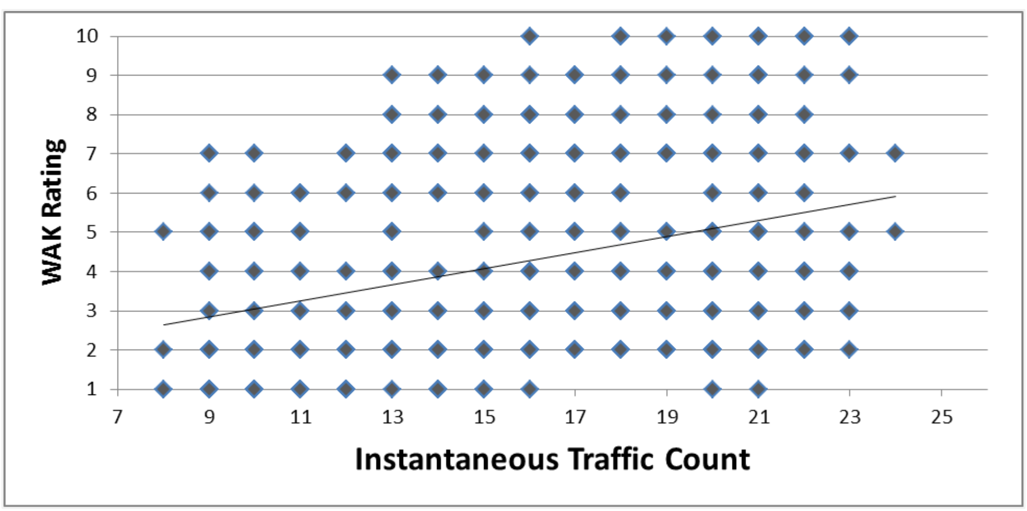

Figure 11. WAK rating as a function of instantaneous traffic count.

The traffic count explained 11.4% of the variance in WAK ratings, and the linear regression model was significant, $p < .001$. The estimates for the linear equations were:

$$\text{WAK Rating} = 0.98 + 0.21 * \text{InstCount} \tag{2}$$

3.2 Functional Near Infra-Red Spectroscopy

We used fnirSoft (Ayaz, 2010) to reduce the fNIRS data. This process included rejecting erroneous optodes that produced erroneous data, filtering motion artifacts and ambient light, and calculating oxygenated and deoxygenated hemoglobin.

Excessive ambient light, motion artifacts, and insufficient contact between the optode and a participant's head all represented unresolvable problems, and we rejected data affected by any of these issues. Any scores above 4,000 mV most likely occurred because the optode was not in contact with the participant's head and was reading ambient light instead. Any value below 400 mV was most likely caused from hair underneath the sensor. We also rejected any optode where the oxygenated and deoxygenated graphs overlapped for the entire length of the experimental run. Besides checking these characteristics from each optode, any optode that showed up as saturated needed to be rejected from the analysis. Most optodes produced usable data at some point, but all did so intermittently. The optode on the top-left temple (Optode 1) was saturated with ambient light on all but three experimental runs. The optode on the top-right temple (Optode 15) was too low to be usable on approximately one-third of the runs. On approximately one-eighth of the experimental runs, the one optode or both of the optodes just above the right eyebrow (Optodes 8 and 10) had inadequate contact with the participant's forehead. Other than these patterns, we had to reject the data from only a few different optodes per experimental run, most of which were because of low scores that probably resulted from hair being underneath an optode.

Upon removal of data from individual optodes, we used a default filter with a sliding window to further subtract motion artifacts and ambient light corruption. The remaining data were then put through a finite impulse response low-pass digital filter that used a Hamming window.

During the first two minutes of each experimental run, the workstation screens were blank, and the participants were not controlling traffic or engaging in any experimental task. We used this time period as a low-workload baseline. During reduction, we set fnirSoft to use these first two minutes of recorded data as the baseline. After we set this baseline, fnirSoft calculated levels of oxygenated and deoxygenated hemoglobin, relative to baseline values, during the rest of each experimental run.

As noted above, all optodes had intermittent issues collecting usable data. Three of the 48 runs (6 per participant) for which we collected the data were not at all usable, and about 67.8% of the individual optode readings were missing across the remaining 45 runs. Therefore, results reported in this section should be treated with caution. Most of the runs in which we had chosen to collect fNIRS data did not run correctly and were not analyzed for many of the analyses in this paper. We therefore conducted exploratory analyses on several subsets of the available data to test the main effect of each of three independent variables for which our usable sample enabled a clean test.

As mentioned in Section 2.8.3, we collected fNIRS data in the same six conditions for all participants. The conditions varied according to CP Location, Algorithm, and Noise. Table 7 summarizes the levels of these variables in each of the available conditions.

Table 7. Conditions for which fNIRS Data were Available, and Levels of Independent Variables

Condition Code	Location	Algorithm	Noise
A-R1	R & RA	Enhancements	No
A-R3	R & RA	Enhancements	No
A-R5	R & RA	Enhancements	No
A-RA5	RA Only	Enhancements	No
B-R1	R & RA	Legacy	Yes
B-R2	R & RA	Enhancements	Yes

To test the effect of Location, we compared the fNIRS data for condition A-RA5—the only RA Only condition for which we collected the data—against the three conditions from which it differed only on CP Location: A-R1, A-R3, and A-R5. These three conditions were equal to each other on all variables in the experiment because of the aforementioned complications with implementing the variations in conformance bounds and likelihood functions. They differed only by the cut times and call sign scrambles that caused every condition/scenario experiment-wide to differ from each other for each individual participant. Therefore, for this analysis, we treated the three conditions as three replications of the same experimental condition. We deemed that a multiple regression analysis with participant as a fixed effect would be sufficiently robust to the resulting unequal sample sizes. Also, this analysis was necessarily exploratory due to the experiment complications and the missing data and we did not intend to draw strong conclusions from the fNIRS results.

For the effect of Algorithm, we compared the fNIRS data for condition B-R1, the only Legacy Algorithm condition for which we collected the data, against the condition from which it differed only on Algorithm: B-R2. The B-R2 run for one participant was one of the completely lost runs, so only seven participants' data could be included in this analysis. To assess Surveillance Noise effects, we compared the fNIRS data for condition B-R2, the only condition with Noise that we could compare cleanly to conditions from which it varied only on Noise presence, against the three conditions with which this variance existed: A-R1, A-R3, and A-R5. Justification for the resulting inequality of sample size is discussed in the preceding paragraph. As with the Algorithm analysis, we were missing one participant's B-R2 data and, therefore, used only seven participants in this analysis.

The available conditions for which we collected fNIRS data did not permit clean assessment of any interactive effects. Therefore we limited our fNIRS analysis to the three exploratory analyses of each main effect.

For each of the three effects, we conducted a multiple regression analysis patterned after the WAK analysis described in Section 3.1. We analyzed the oxygenated hemoglobin level relative to the run-specific baseline as the dependent measure (Ayaz, Willems et al., 2010). The models used the mean across all valid optode readings for the 240 observations in each 2-minute period as a unit of observation. We used the maximum aircraft count in the sector being worked over the corresponding 2-minute period, and the number of minutes into the run at which the interval began (time-on-task) as covariates. We entered these variables, along with the independent variable of interest for the analysis, into the regression equation as fixed effects, and participant as a random effect, to create a within-subjects model. For fNIRS data, we did not consider using a multilevel approach with unique run nested within participant, because the data entered into the analysis were already corrected for a run-specific baseline.

Location did not have a significant effect on the oxygenated hemoglobin level. However, the Algorithm variable significantly affected this measure. The levels in the Legacy algorithm conditions ($M = 1.94$, $SD = 1.82$) were higher than they were in the algorithmic enhancements ($M = 1.50$, $SD = 1.26$), indicating a reduction in mental workload with the algorithmic improvements. Figure 12 shows this difference.

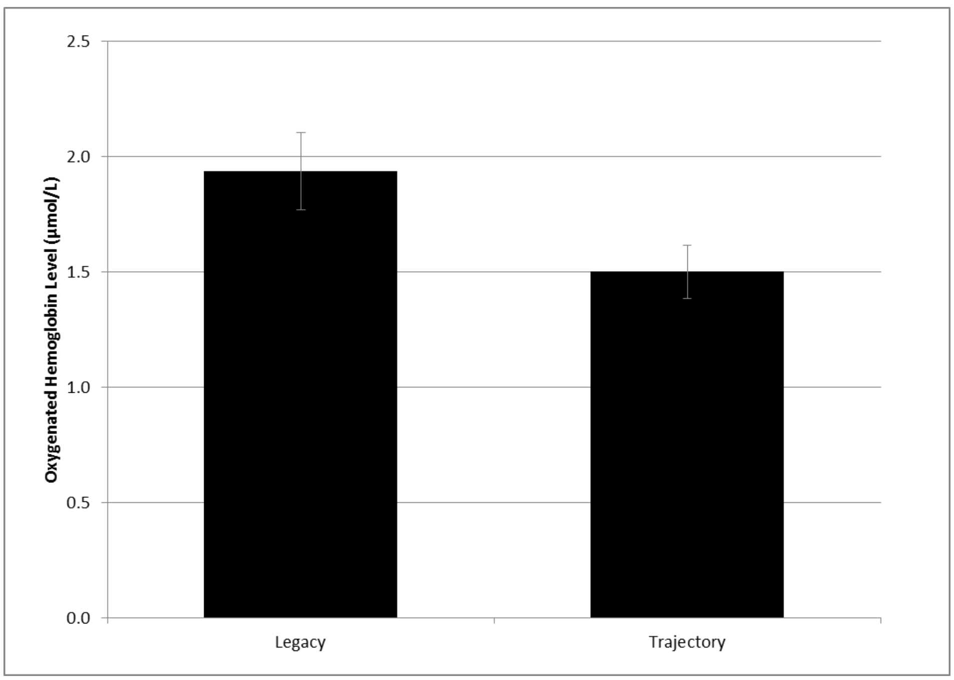

Figure 12. Mean oxygenated hemoglobin levels from fNIRS, by Algorithm condition.

The result of our regression analysis showed a statistically significant effect of Algorithm, $F(1, 229.6) = 4.07$, $p < .05$.

Our analysis also revealed a trend toward an effect of Surveillance Noise. The oxygenated hemoglobin levels, and therefore presumably the mental workload, were higher in the Noise conditions ($M = 1.50$, $SD = 1.26$) than in the No-Noise conditions ($M = 1.44$, $SD = 1.79$). The trend toward the effect of Noise on the fNIRS data is depicted in Figure 13.

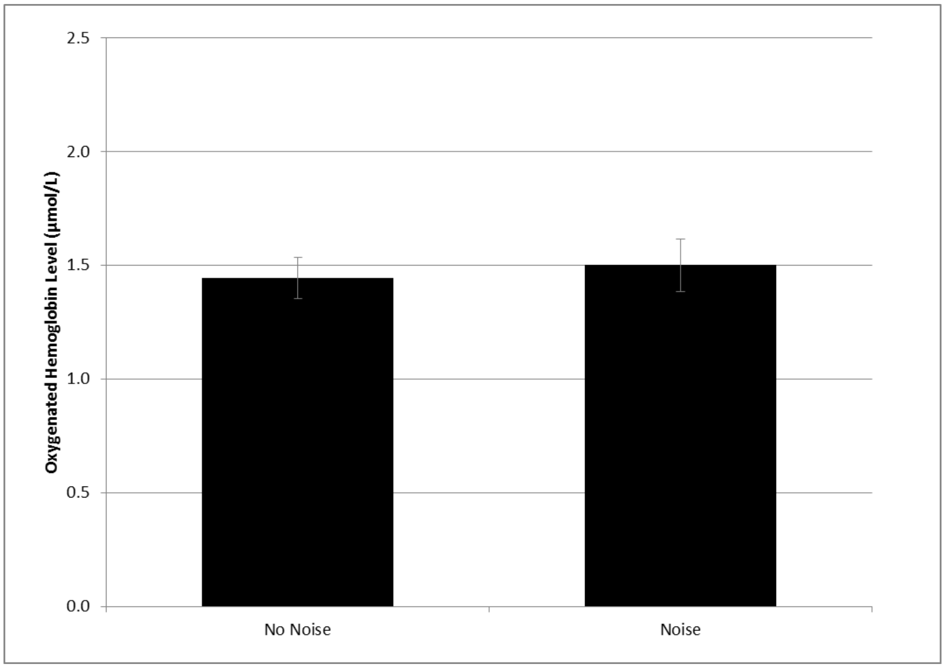

Figure 13. Mean oxygenated hemoglobin levels from fNIRS, by Noise condition.

The result of our regression analysis showed a trend toward statistical significance for Noise. Detailed results of fNIRS analysis are found in Appendix I.

3.3 Subjective Assessments

We conducted the analysis of the PSQ and Over-The-Shoulder Rating Form data by using Bayesian ANOVA, which provided several advantages. First, this method is robust against non-normal distributions. Second, this method allowed us to analyze a Region of Practical Equivalence (ROPE), which provides us with a range of parameters within which we conclude that there are no practical differences between two conditions (Kruschke, 2015). To conduct the Bayesian analysis, we centered the data at zero and normalized the distribution. We reported a 95% High-Density Interval (HDI), which provides an estimate of possible differences between conditions, and any interval that does not include zero allowed us to reject the null hypothesis. We also reported a direct comparison to zero that provides an estimate of credible difference above and below zero, and we view this as a liberal estimate that is more analogous to a one-tailed test. To determine the ROPE, we considered the fact that the questionnaires only allowed controllers to report whole integers (e.g., 1, 2, 3, and so on.) and set critical values to 0.5, which is halfway between any reportable value. In a manner consistent with the normalization of the distribution of outcomes, we used Formula 2 to transform this value and provide a unique range for each outcome.

$$ROPE = 0 \pm (0.5 \div SD) \tag{3}$$

The following subsections report only significant differences and meaningful findings with the ROPE. For a more detailed report of all findings, see Appendix J.

Bayesian analysis is useful for working with non-normal data, but there are differences with classical statistics that use a mean and standard deviation. Bayesian analysis uses the R-package for Just Another Gibbs Sampler (JAGS; cf. Plummer, 2003) to create a simulation that provides a distribution of credible values for the conditions in this analysis. When findings are significant, we report the mean from the simulation's estimate of this distribution. Also, this simulation does not provide estimates of standard error, and we do not report them in this analysis. Instead, the simulation provides an estimated distribution of possible standard deviations, and Appendix J reports the mean from this distribution of standard deviations.

3.3.1 Post-Scenario Questionnaire

3.3.1.1 Algorithm by CP Location Analysis

The analysis of nuisance alerts indicated that controllers reported more nuisance alerts with the legacy CP than with the algorithmic enhancements, with an HDI interval ranging from 0.185 to 0.862 (see Figure 14 for ROPE and comparison to zero). With an HDI ranging from 0.339 to 1.714, controllers reported more nuisance alerts when the CP was on both the R- and RA-Side than the RA-Side alone (see Figure 15 for ROPE and comparison to zero). This relationship is best explained by the interaction between these variables, which had an HDI ranging from 0.388 to 1.714 (see Figure 16 for ROPE and comparison to zero). Figure 17 illustrates this relationship, where these main effects were attributed to controllers only reporting a high number of nuisance alerts for an R-Side probe that used legacy algorithms. For the R-Side position with algorithmic enhancements or any condition on the RA-Side Only, controllers did not report many, if any, nuisance alerts.

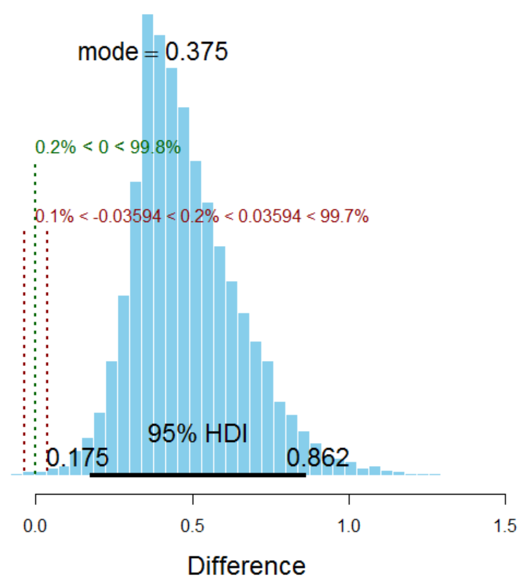

Figure 14. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of nuisance alerts.

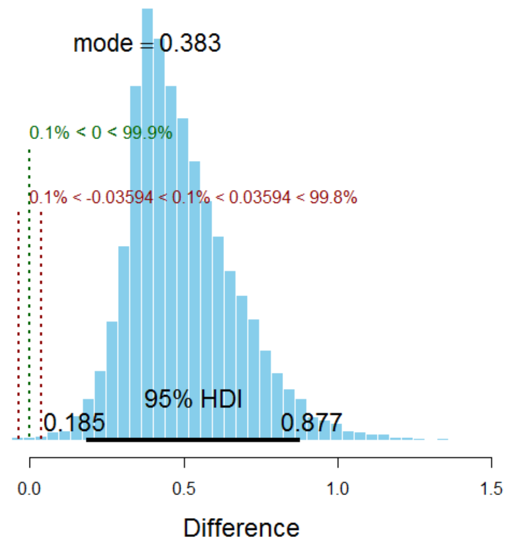

Figure 15. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Algorithm analysis of nuisance alerts.

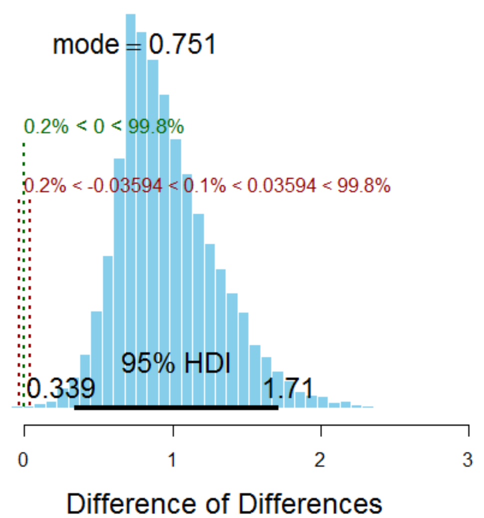

Figure 16. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location by CP Algorithm analysis of nuisance alerts.

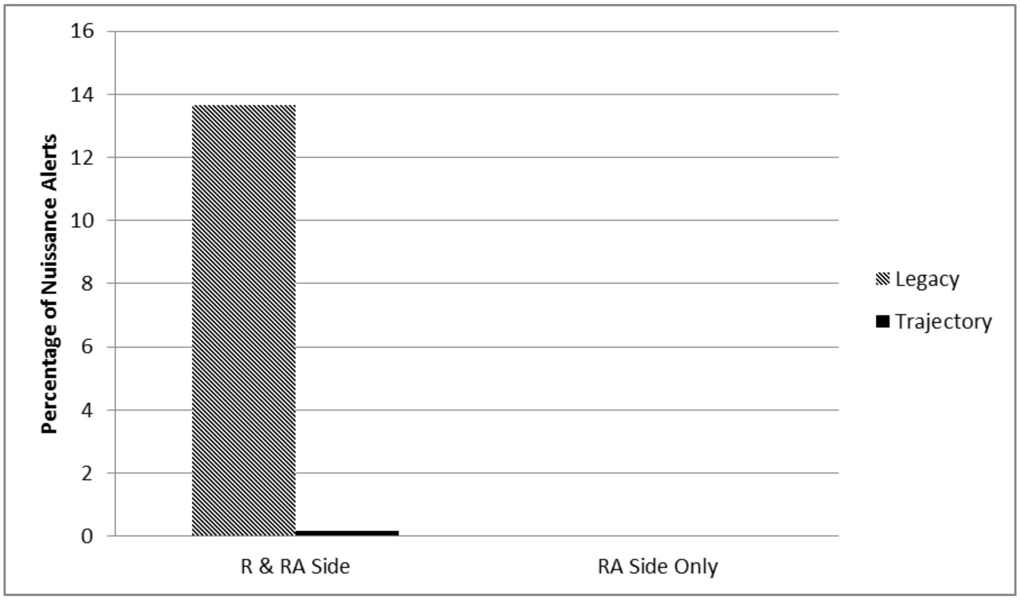

Figure 17. Graph of the mean estimates from the CP Location by CP Algorithm analysis of the percentage of nuisance alerts.

When asked to rate usefulness, controllers rated the CP higher when it was on both the R- and RA-Side. The HDI for this relationship ranged from 0.063 to 1.773 (see Figure 18 for ROPE and comparison to zero). This range corresponds to rating the CP between 0.181 and 5.094 points (with an average of 2.594) higher when the CP was on the R- and RA-Side (see Figure 19).

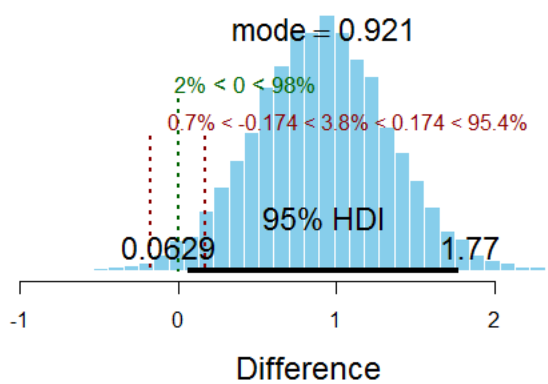

Figure 18. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of usefulness ratings.

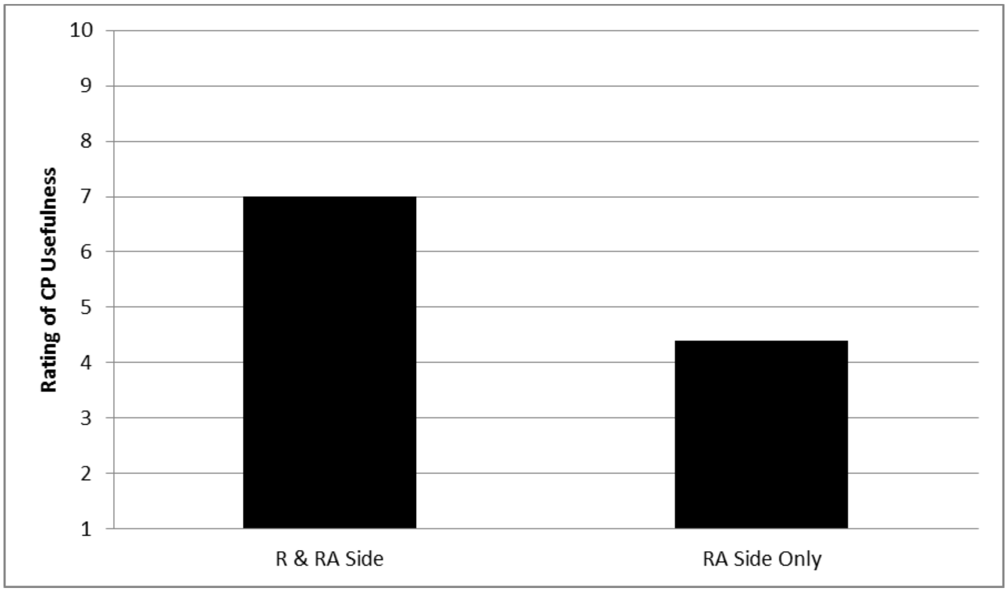

Figure 19. Graph of the mean estimates from the CP Location analysis of the CP usefulness ratings.

When asked about the CP distracting controllers, an HDI ranging from 0.044 to 1.401 indicated that controllers rated the CP as being more distracting with legacy algorithms than with algorithmic enhancements (see Figure 20 for ROPE and comparison to zero). With HDI ranging from 0.014 to 1.377 (see Figure 21 for ROPE and comparison to zero), controllers rated the CP as being more distracting on R- and RA-Side than on the RA-Side alone. While the HDI for the interaction was not significant (ranging from -0.062 to 2.684), 97.2% of credible values were above zero (see Figure 22). Similar to the pattern with nuisance alerts, controllers rated the distraction highest when the R-Side CP used algorithmic enhancements (see Figure 23).

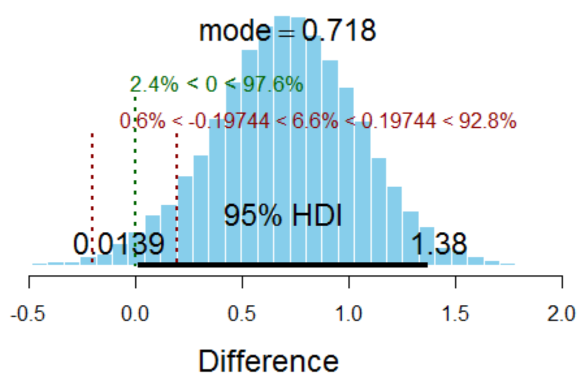

Figure 20. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of distraction ratings.

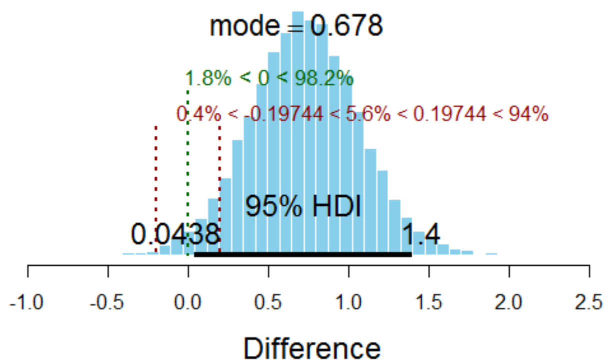

Figure 21. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Algorithm analysis of distraction ratings.

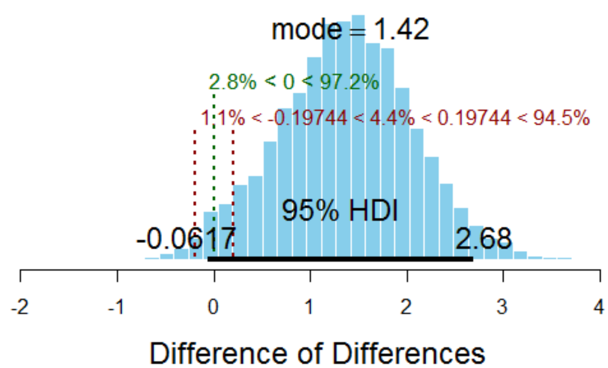

Figure 22. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location by CP Algorithm analysis of distraction ratings.

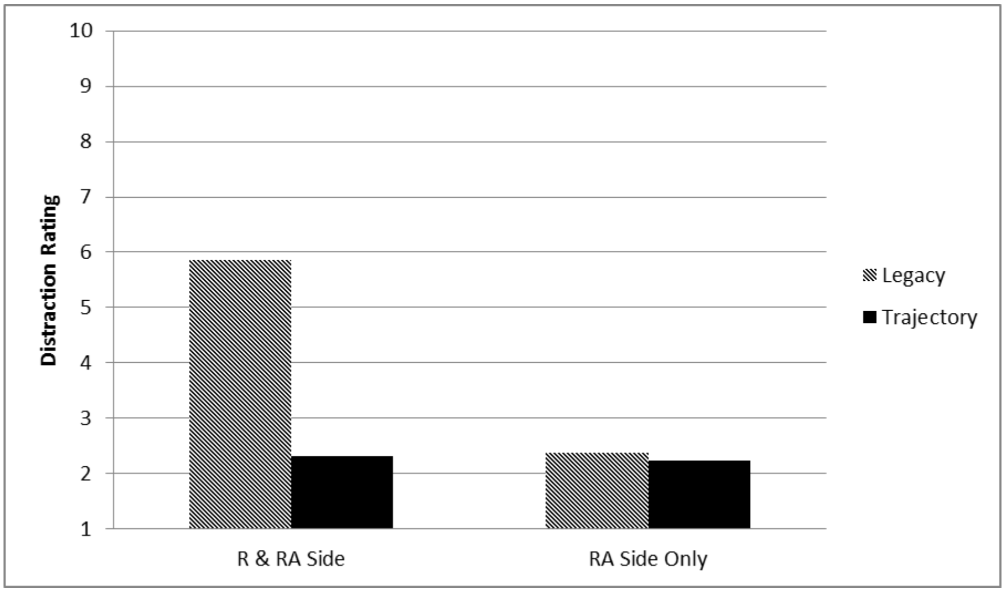

Figure 23. Graph of the mean estimates from the CP Location by CP Algorithm analysis of distraction ratings.

When responding to the question about whether they were prepared to trust the CP, both the HDI and a comparison to zero indicated that there were no differences between any of the conditions. With regard to the ROPE, the analysis of CP Location (see Figure 24), CP Algorithm (see Figure 25), and the interaction (see Figure 26) had 99.64%, 99.87%, and 97.25 % of credible values within the ROPE, indicating that we could consider responses practically equivalent. Furthermore, estimates of the mean (0.981), median (0.997), and mode (0.999) indicated that controllers said that they trusted the CP (coded as 1.00).

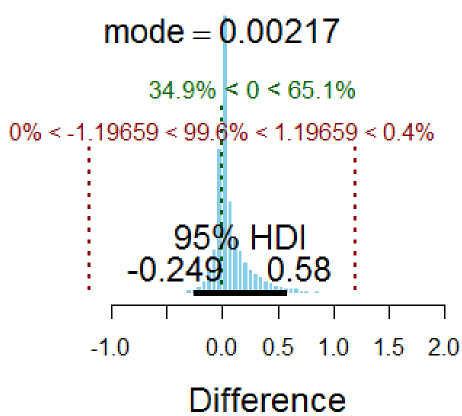

Figure 24. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of trust ratings.

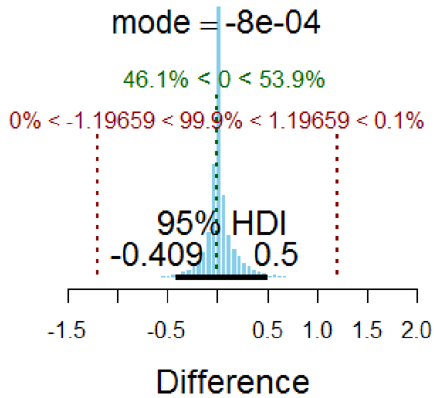

Figure 25. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Algorithm analysis of trust ratings.

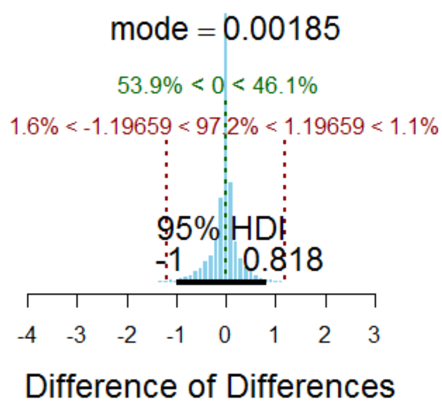

Figure 26. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location by CP Algorithm analysis of trust ratings.

At the end of the PSQ, there were a series of strategy statements (see Appendix C and Table J1) that highlight differences according to the CP Location. For the statement "I responded to a conflict without question after I received an alert from the Conflict Probe," 97.4% of credible values were above zero (the HDI ranged from -0.033 to 1.530), indicating that controller ratings of agreement were higher when the CP was on the R- and RA-Side than the RA-Side alone (see Figure 27 and Figure 28). For the statement "I noticed alerts from the Conflict Probe and believed that it was correct, but I checked other data before I responded to the conflicts," 96.0% of credible values were above zero (the HDI ranged from -0.079 to 1.454), indicating that controllers were more likely to agree when the CP was on the R- and RA-Side than the RA-Side alone (see Figure 29 and Figure 30). For the statement "I noticed alerts from the Conflict Probe and did not know if they were correct, so I checked other data before responding to the conflicts," 97.9% of credible values were above zero (the HDI ranged from 0.026 to 1.614), indicating that controllers were more likely to agree when the CP was on the R- and RA-Side than the RA-Side Only (see Figure 31 and Figure 32). This pattern reversed for the statement "I detected and resolved conflicts without using the Conflict

Probe," which had 98.9% of credible values below zero (the HDI ranged from -1.528 to -0.134), indicating that controllers were more likely to agree when the CP was on the RA-Side Only than on the R- and RA-Side (see Figure 33 and Figure 34).

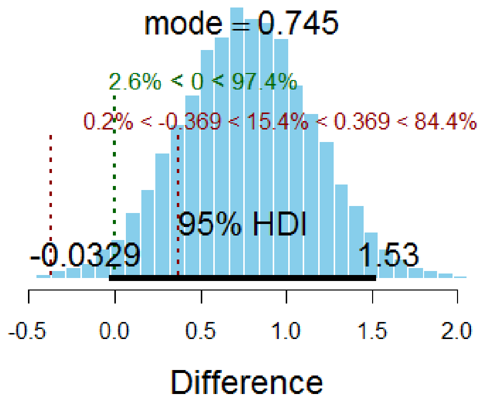

Figure 27. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of the strategy statement to respond to CP Alert without question.

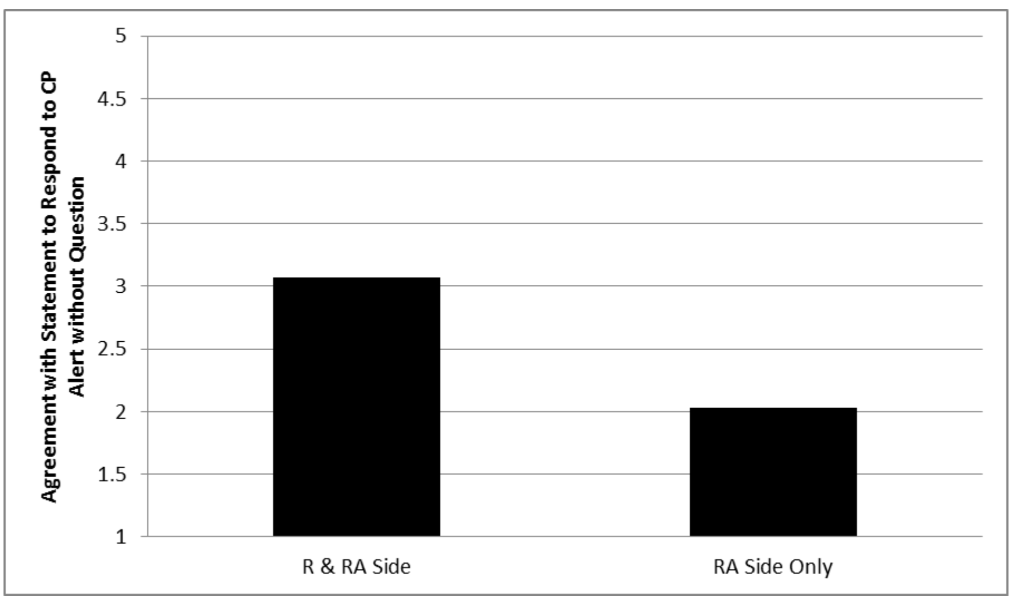

Figure 28. Graph of the mean estimates from the CP Location analysis of the strategy statement to respond to CP Alert without question (1 = Never; 5 = Always).

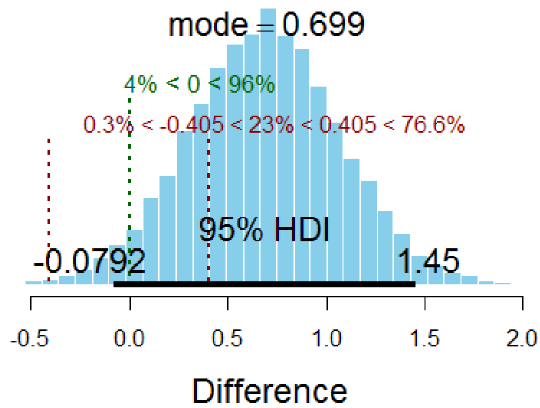

Figure 29. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of the strategy statement to believe CP Alert but validate with other data.

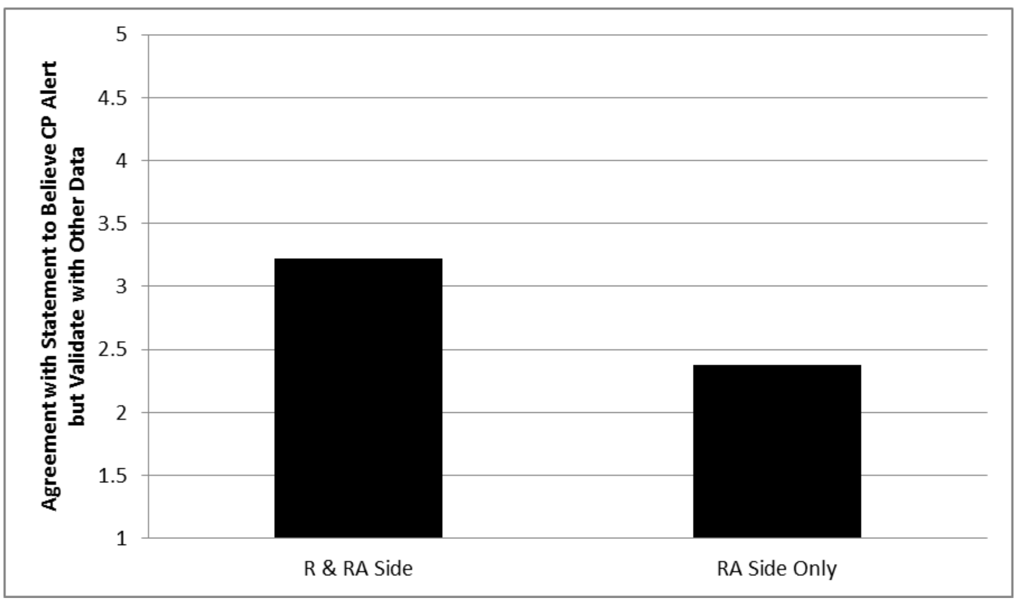

Figure 30. Graph of the mean estimates from the CP Location analysis of the strategy statement to believe CP Alert but validate with other data (1 = Never; 5 = Always).

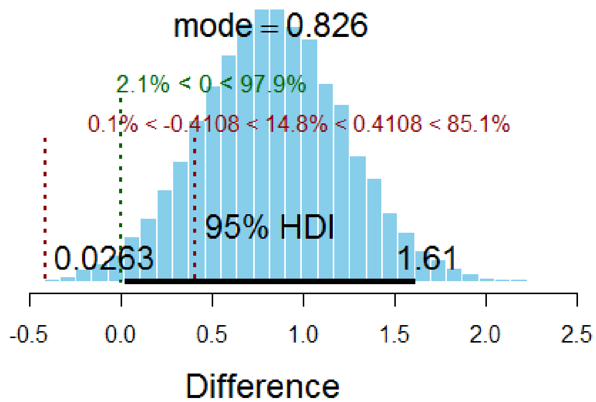

Figure 31. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of the strategy statement to not know accuracy of CP Alert and validate with other data.

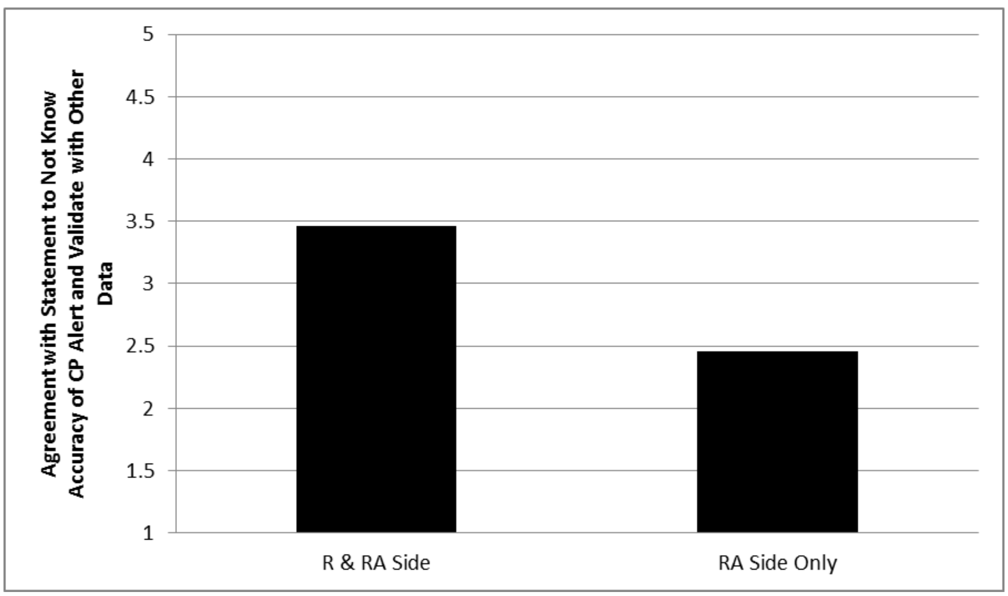

Figure 32. Graph of the mean estimates from the CP Location analysis of the strategy statement to not know accuracy of CP Alert and validate with other data (1 = Never; 5 = Always).

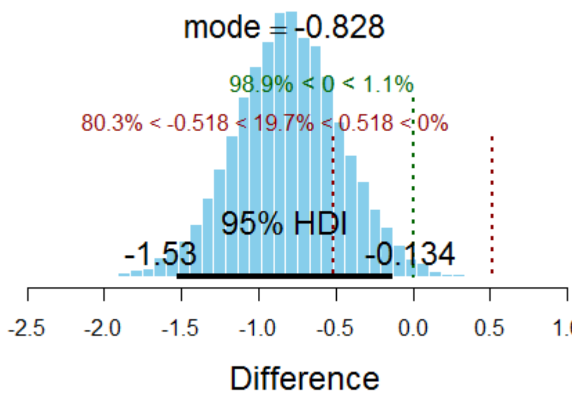

Figure 33. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of the strategy statement to detect and resolve conflicts without the CP.

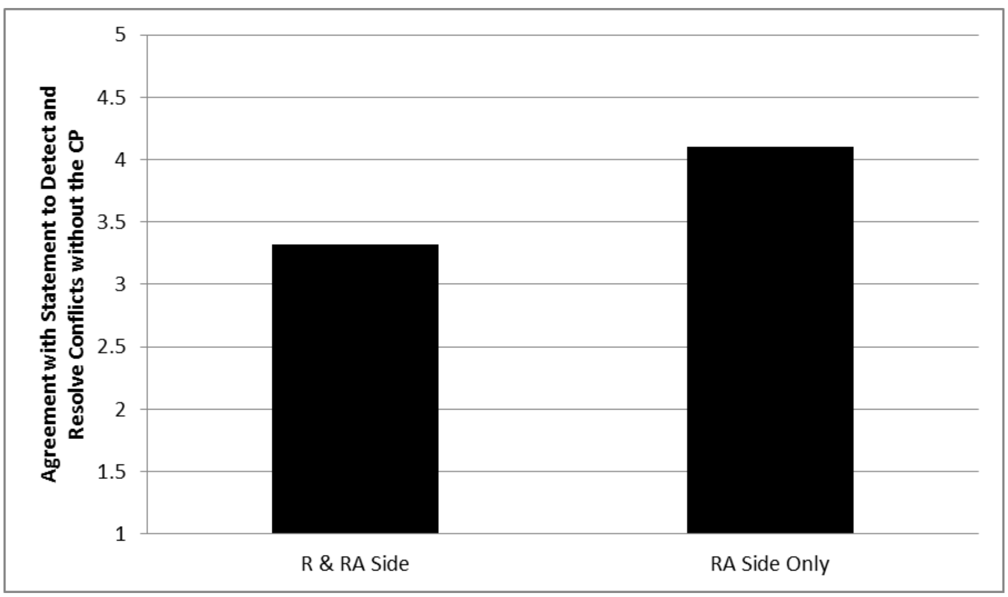

Figure 34. CP Location analysis of the strategy statement to detect and resolve conflicts without the CP (1 = Never; 5 = Always).

3.3.1.2 Radar Surveillance Noise by CP Location Analysis

For CP usefulness ratings, an HDI ranging from 0.813 to 2.118 (see Figure 35 for ROPE and comparison to zero) indicated that there was an effect of CP Location. Figure 36 illustrates that controllers rated the CP more useful when it was on the R- and RA-Side than the RA-Side Only. There was no effect involving radar surveillance noise.

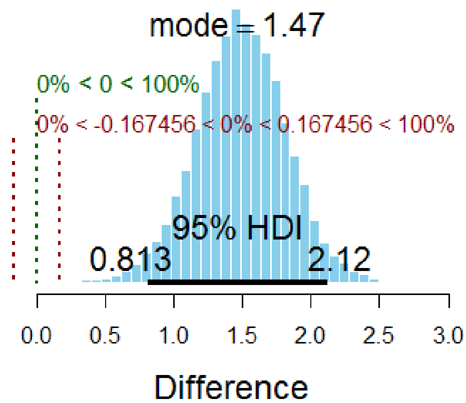

Figure 35. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of usefulness ratings.

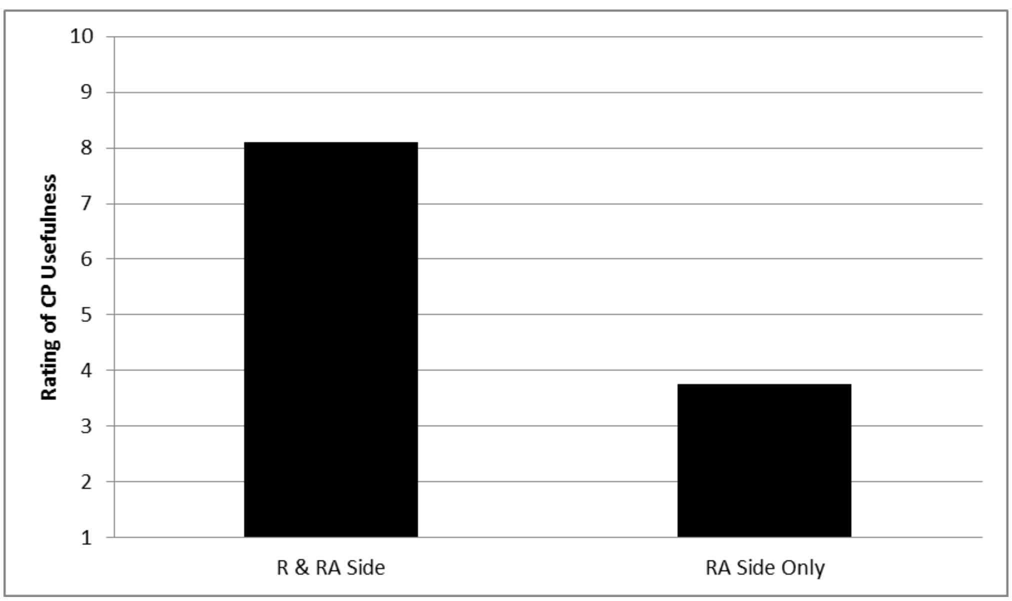

Figure 36. Graph of the mean estimates from the CP Location analysis of the CP usefulness ratings.

The PSQ had a question asking about confidence in the detection accuracy of the CP. With 97.18% of credible values above zero (the HDI ranged from -0.024 to 1.625), controllers said that they had more confidence in the detection accuracy of the CP when it was on the R- and RA-Side than the RA-Side Only (see Figure 37 and Figure 38).

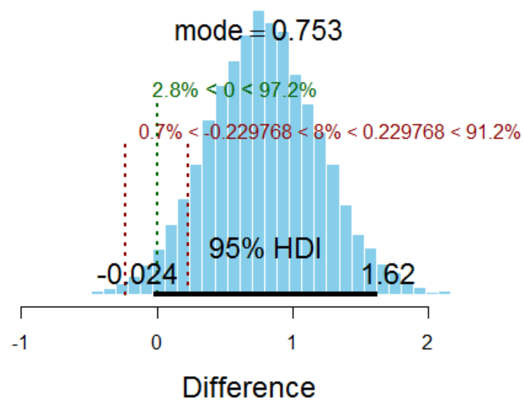

Figure 37. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of controller confidence in CP accuracy.

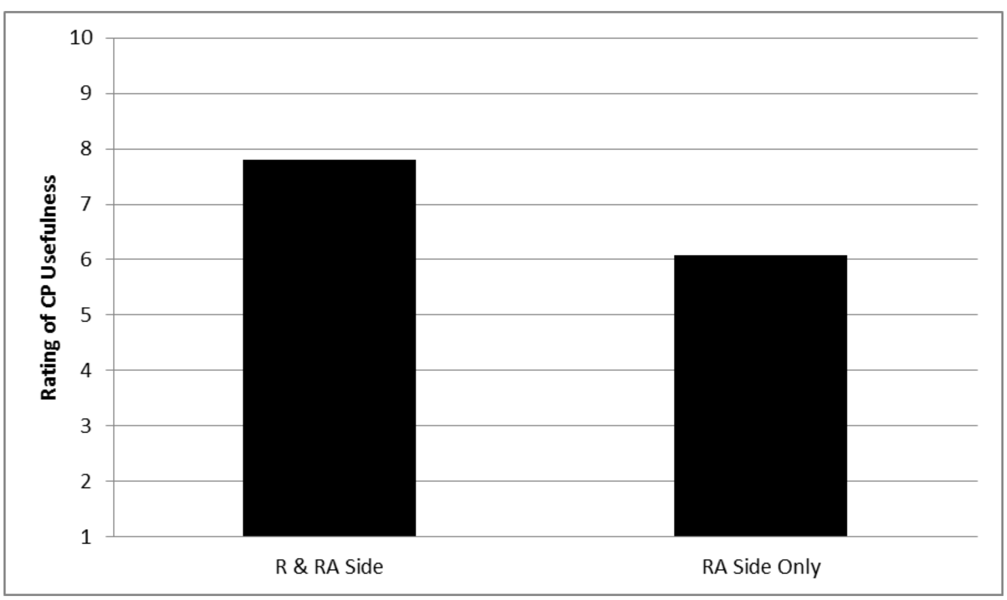

Figure 38. Graph of the mean estimates from the CP Location analysis of controller confidence in CP accuracy.

At the end of the PSQ, there were a series of strategy statements (see Appendix C as well as Appendix J; Table J2) that highlight differences according to the CP Location. For the statement "I responded to a conflict without question after I received an alert from the Conflict Probe," 99.9% of credible values were above zero (the HDI ranged from 0.436 to 1.856), indicating that controllers were more likely to agree when the CP was on the R- and RA-Side than the RA-Side Only (see Figure 39 and Figure 40). For the statement "I noticed alerts from the Conflict Probe and believed that it was correct, but I checked other data before I responded to the conflicts," 97.487% of credible values were above zero (the HDI ranged from -0.007 to 1.682), indicating that controllers were more likely to agree when the CP was on the R- and RA-Side than the RA-Side Only (see Figure 41 and Figure 42). This pattern reversed for the statement "I detected and resolved conflicts without using the Conflict Probe," which had 99.9% of credible values below zero (the HDI ranged from -1.667 to -0.204), indicating that controllers were more likely to agree when the CP was on the RA-Side Only than the R- and RA-Side (see Figure 43 and Figure 44).

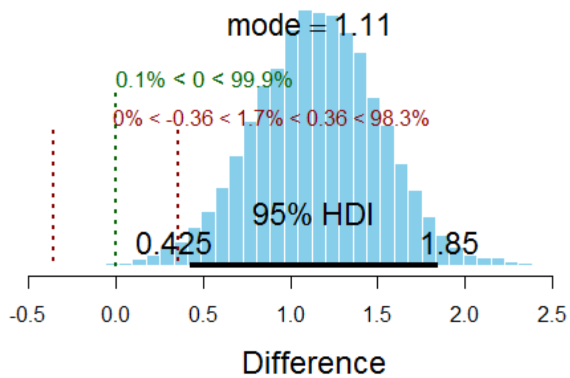

Figure 39. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of the strategy statement to respond to CP Alert without question.

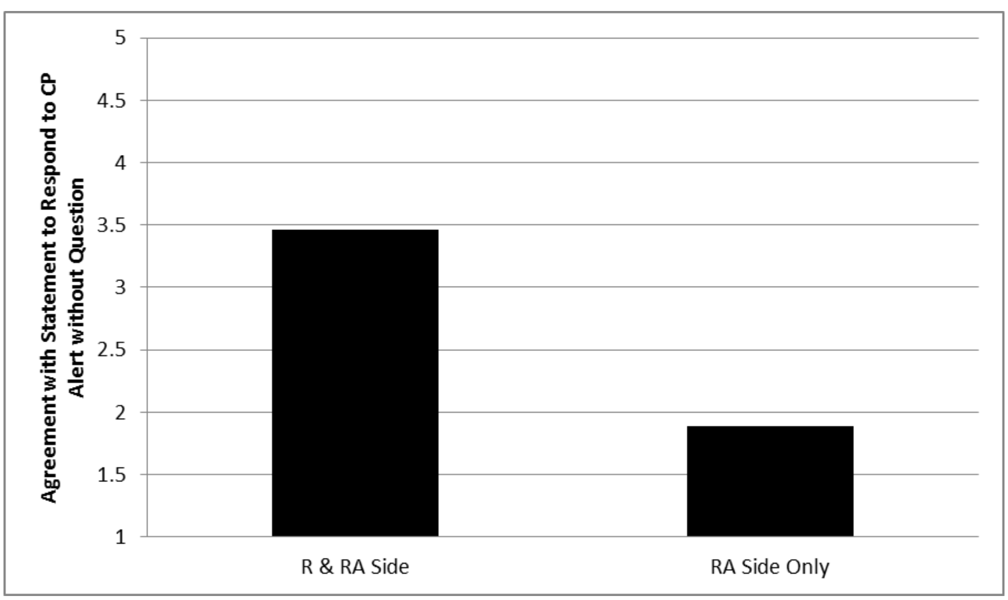

Figure 40. Graph of the mean estimates from the CP Location analysis of the strategy statement to respond to CP Alert without question (1 = Never; 5 = Always).

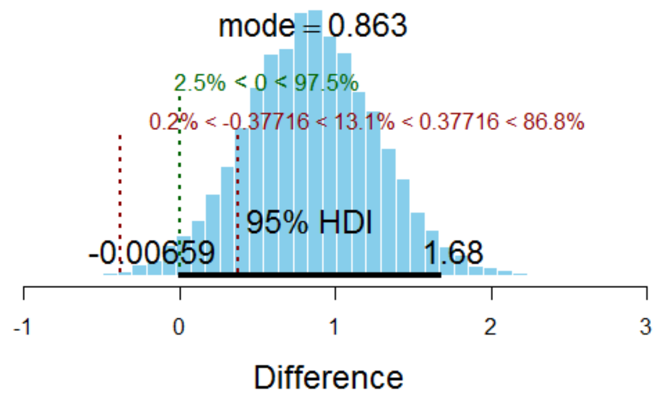

Figure 41. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of the strategy statement to believe CP Alert but validate with other data.

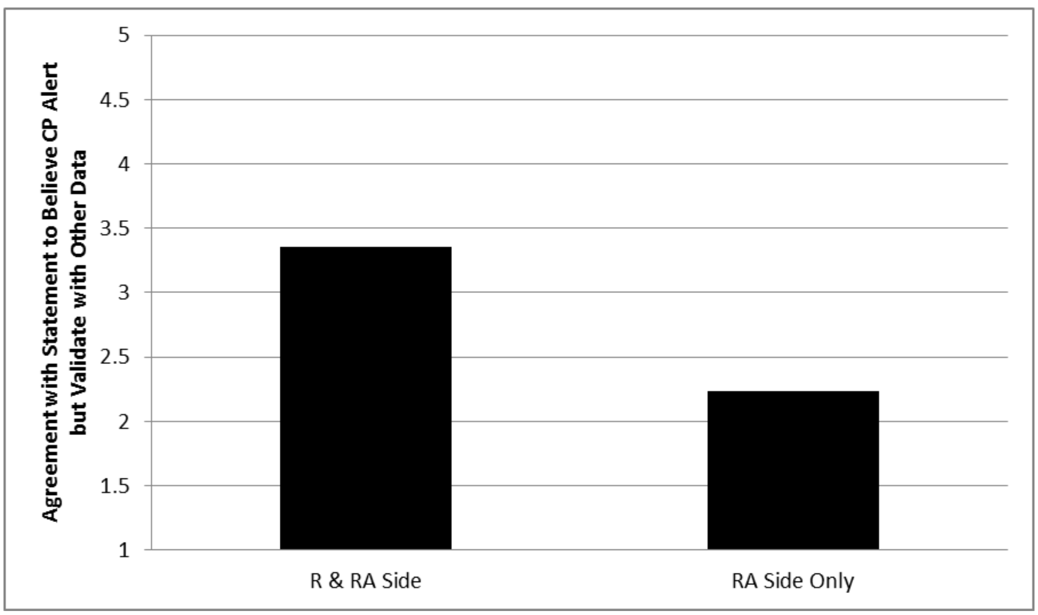

Figure 42. Graph of the mean estimates from the CP Location analysis of the strategy statement to believe CP Alert but validate with other data (1 = Never; 5 = Always).

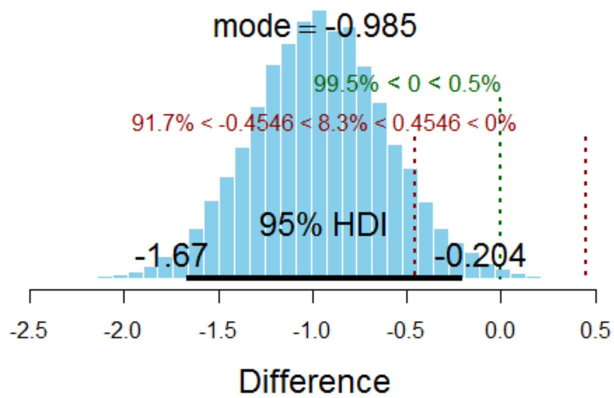

Figure 43. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for CP Location analysis of the strategy statement to detect and resolve conflicts without the CP.

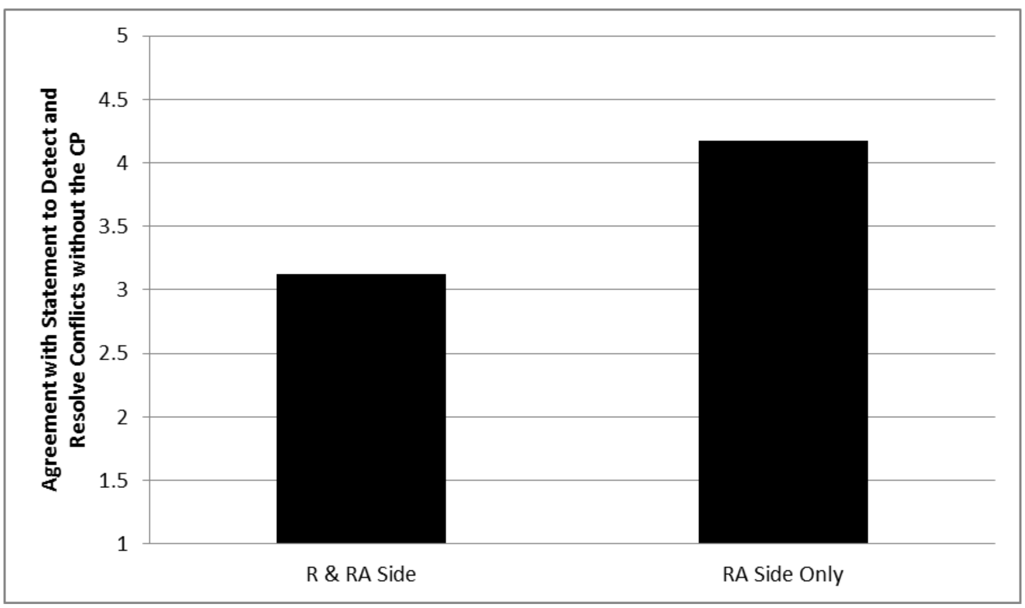

Figure 44. Graph of the mean estimates from the CP Location analysis of the strategy statement to detect and resolve conflicts without the CP (1 = Never; 5 = Always).

3.3.2 Over-The-Shoulder Rating Form

The following subsections present analyses of the ratings that the over-the-shoulder observers provided during the study.

3.3.2.1 Algorithm by CP Location Analysis

With the over-the-shoulder rating form, observers often gave controllers consistently high ratings on a number of different outcomes. There was no effect of the manipulations for these outcomes, but a large percentage of credible responses were within the ROPE. To simplify this reporting, we summarize these results in Table 8 with a list of questions and findings per condition.

Table 8. Summary of Over-the-Shoulder Ratings with a Significant Percentage of Credible Values within the ROPE for CP Algorithm by CP Location Analysis

	% Credible Values within ROPE		
Over-The-Shoulder Rating Item	CP Location	Algorithm	CP Location by Algorithm
2. Maintaining Safe and Efficient Traffic Flow: Sequencing Arrival and Departure Aircraft Efficiently	99.2%	98.9%	89.5%
3. Maintaining Safe and Efficient Traffic Flow: Using Control Instructions Effectively	100%	100%	99.3%
6. Maintaining Attention and Situation Awareness: Ensuring Positive Control	90.1%	95.3%	83.9%
7. Maintaining Attention and Situation Awareness: Detecting Pilot Deviations from Control Instructions	82.5%	95.7%	81.2%
14. Providing Control Information: Providing Essential ATC Information	100%	100%	100%
15. Providing Control Information: Providing Additional ATC Information	100%	100%	100%
16. Providing Control Information: Overall Providing Control Information	100%	100%	100%
17. Technical Knowledge: Showing Knowledge of LOAs and SOPs	99.9%	99.8%	98.4%
18. Technical Knowledge: Showing Knowledge of Aircraft Capabilities and Limitations	100%	100%	100%
19. Technical Knowledge: Overall Technical Knowledge	100%	100%	99.8%
20. Voice Communications: Using Proper Phraseology	100%	100%	100%
21. Voice Communications: Communicating Clearly and Efficiently	100%	100%	100%
22. Voice Communications: Listening for Pilot Read backs and Requests	100%	100%	100%
23. Voice Communications: Overall Voice Communication	100%	100%	100%

For maintaining attention and situation awareness of aircraft positions, an HDI ranging from -1.490 to -0.006 (see Figure 45 for ROPE and comparison to zero) indicated that CP Location impacted observer ratings. Figure 46 illustrates that observers rated controller attention and awareness of aircraft positions higher when the CP was only on the RA-Side than the R- and RA-Side.

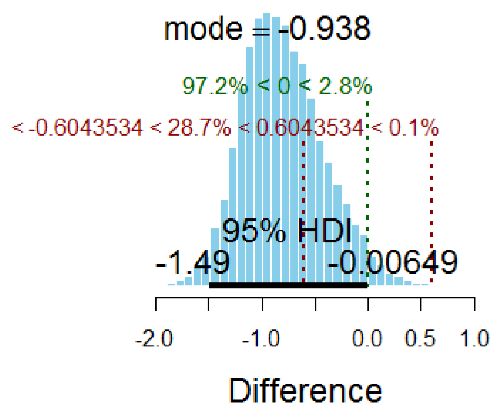

Figure 45. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of maintaining attention and situation awareness of aircraft positions.

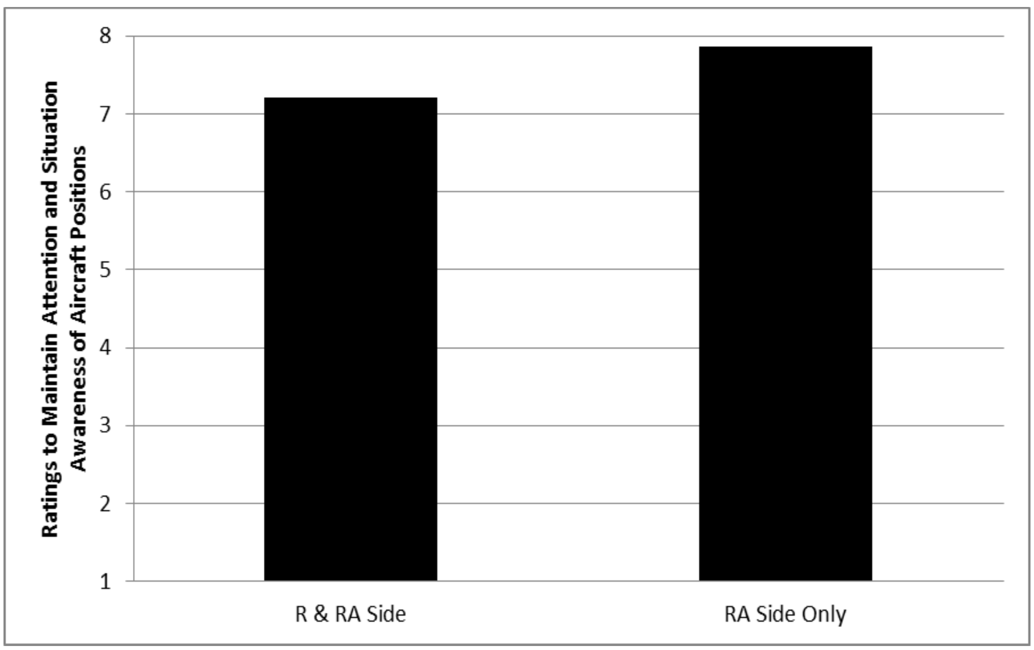

Figure 46. Graph of the mean estimates from the CP Location analysis of ratings of maintaining attention and situation awareness of aircraft positions.

For maintaining overall attention and situation awareness, 95.8% of credible values (see Figure 47) were below zero (the HDI ranged from -1.583 to 0.058), indicating that CP Location impacted observer ratings. Figure 48 illustrates that observers rated overall attention and situation awareness higher when the CP was only on the RA-Side than when it was on the R- and RA-Side.

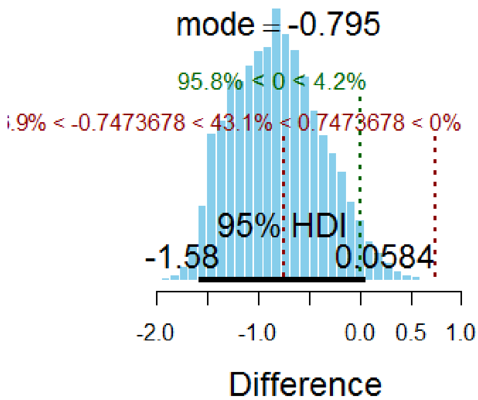

Figure 47. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of maintaining overall attention and situation awareness.

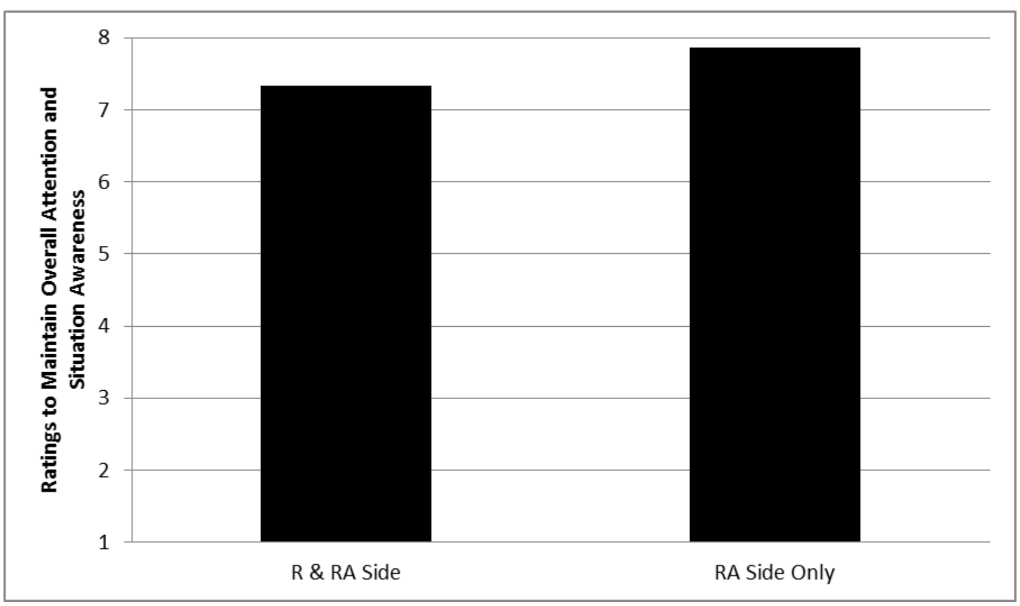

Figure 48. Graph of the mean estimates from the CP Location analysis of ratings of maintaining overall attention and situation awareness.

For questions about prioritizing items, observer ratings indicated that there was an effect of the manipulations on preplanning control actions. For CP Location (see Figure 49 for ROPE and comparison to zero), an HDI ranging from -0.915 to -0.079 indicated that ratings were higher when the probe was on the RA-Side Only. For algorithm (see Figure 50 for ROPE and comparison to zero), an HDI ranging from 0.186 to 1.092 indicated that ratings were higher with the legacy algorithm. An HDI ranging from 0.154 to 1.825 indicated that there was an interaction between CP Location and CP Algorithms (see Figure 51 for ROPE and comparison to zero). Figure 52 illustrates that observer ratings declined only for the R-Side display of alerts with algorithmic enhancements.

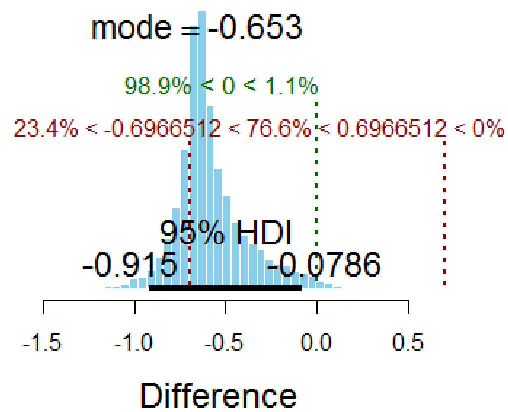

Figure 49. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of prioritizing/preplanning control actions.

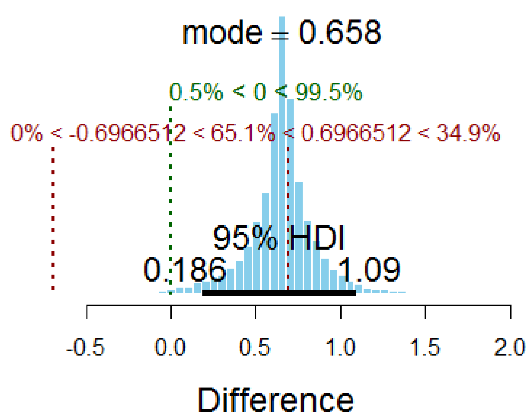

Figure 50. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Algorithm analysis of ratings of prioritizing/preplanning control actions.

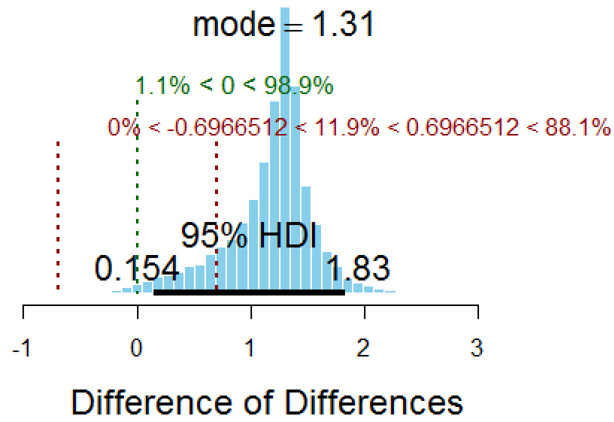

Figure 51. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location by CP Algorithm analysis of ratings of prioritizing/preplanning control actions.

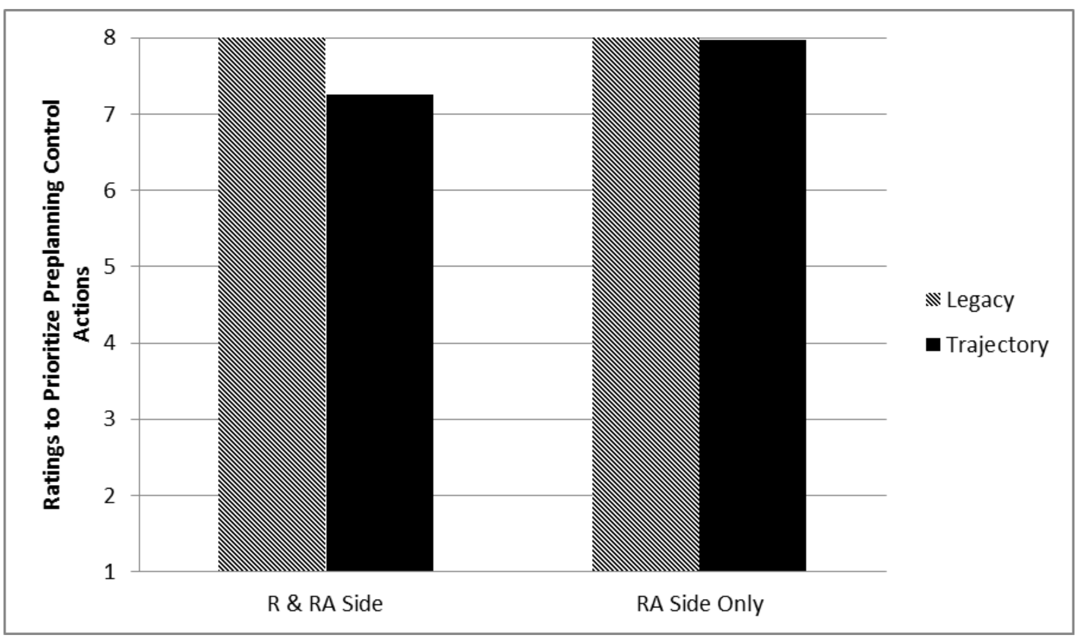

Figure 52. Graph of the mean estimates from the CP Location by CP Algorithm analysis of ratings of prioritizing/preplanning control actions.

For prioritizing the handling control tasks for several aircraft, 96.7% of credible values (the HDI ranged from -0.083 to 1.448), indicating that CP Algorithm influenced observer ratings (see Figure 53). Figure 54 illustrates that observers rated the controller's prioritization higher when they were using legacy algorithms than with algorithmic enhancements.

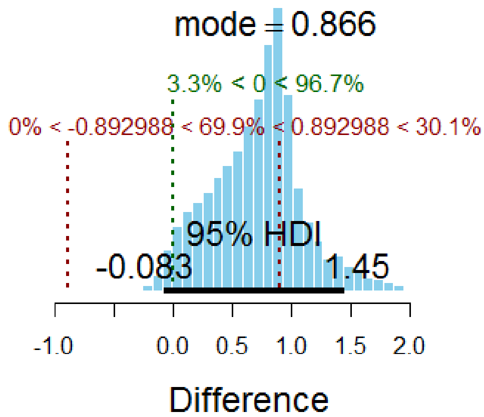

Figure 53. Posterior distribution displaying a 95% HDI, comparison to zero and ROPE for the CP Algorithm analysis of ratings of prioritizing/handling control tasks for several aircraft.

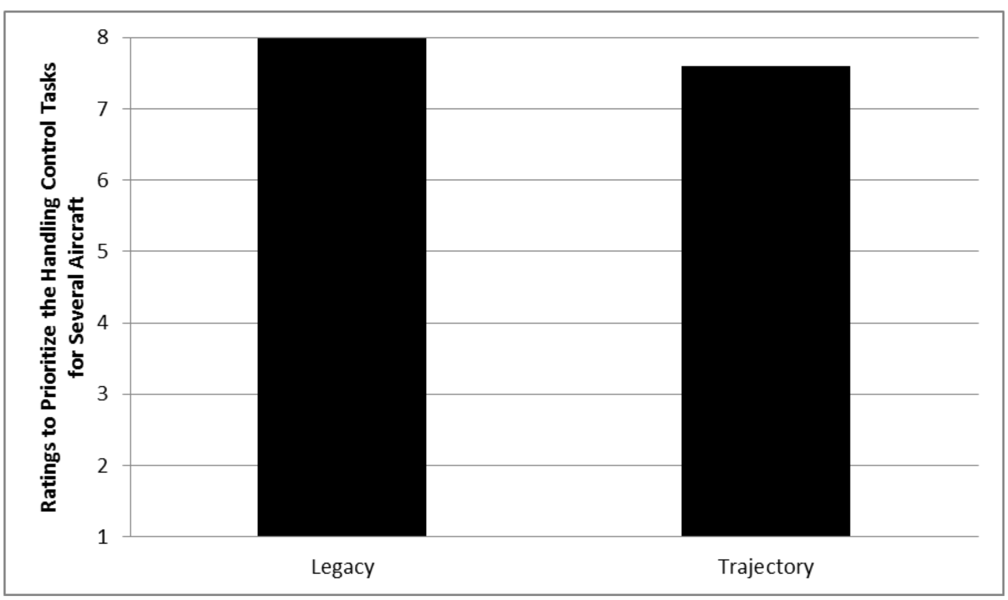

Figure 54. Graph of the mean estimates from the CP Algorithm analysis of ratings of prioritizing/handling control tasks for several aircraft.

There was a question about using the CP in appropriate situations, and results indicated that the manipulations had an impact on observer ratings. An HDI ranging from 0.398 to 1.692 (see Figure 55 for ROPE and comparison to zero) indicated that there was an effect of CP Location. Figure 56 illustrates that ratings were higher when the CP was on the R- and RA-Side. When compared to zero, 97.1% of credible values were above zero (HDI ranged from -0.045 to 1.129), indicating that CP Algorithm had a weak effect on observer ratings (see Figure 57). Figure 58 illustrates that ratings were higher for the legacy algorithm.

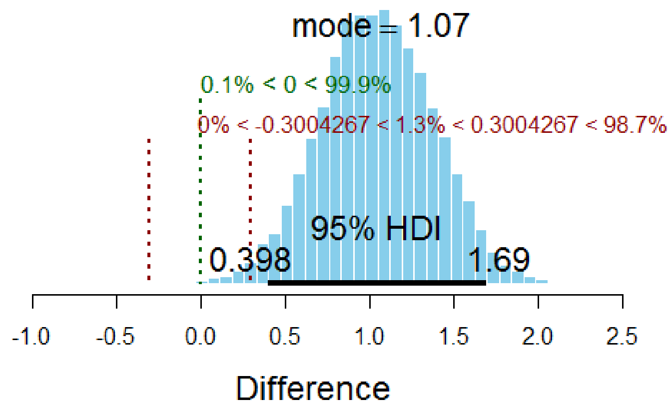

Figure 55. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of using the CP in appropriate situations.

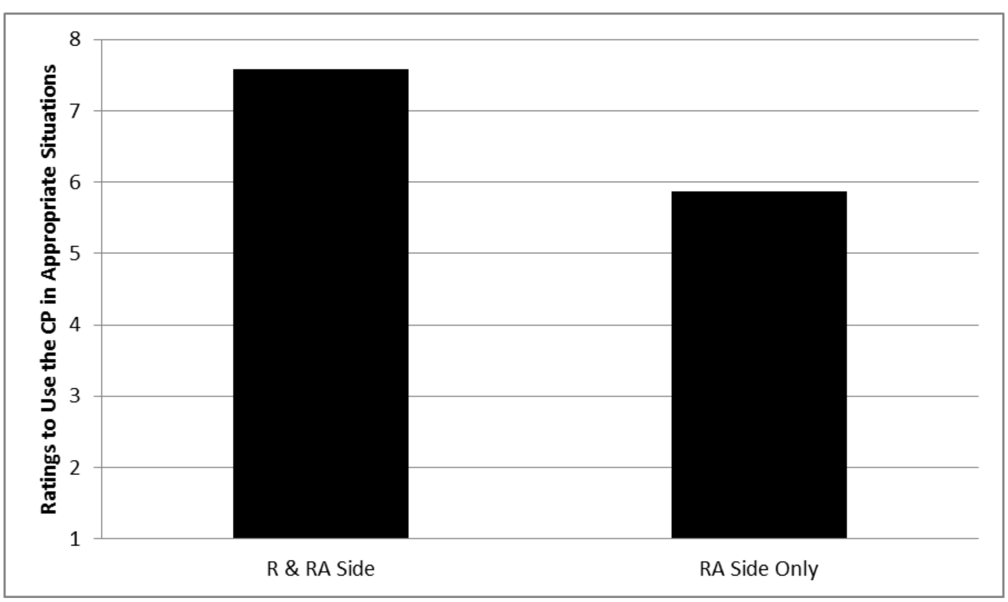

Figure 56. Graph of the mean estimates from the CP Location analysis of ratings of using the CP in appropriate situations.

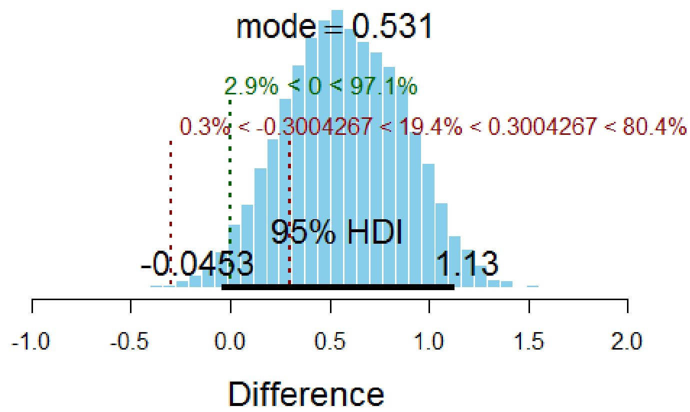

Figure 57. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Algorithm analysis of ratings of using the CP in appropriate situations.

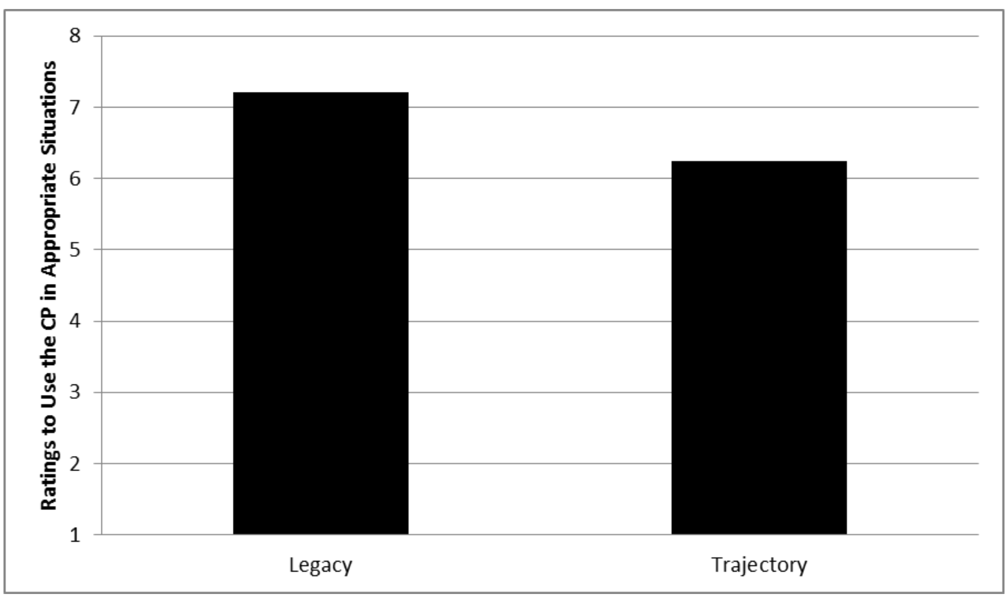

Figure 58. Graph of the mean estimates from the CP Algorithm analysis of ratings of using the CP in appropriate situations.

Observers rated controllers on their ability to take full advantage of the CP in a timely and efficient manner. An HDI ranging from 0.092 to 1.691 indicated that CP Location had an impact on ratings (see Figure 59 for ROPE and comparison to zero). Figure 60 illustrates that ratings were higher when the CP was on the R- and RA-Side.

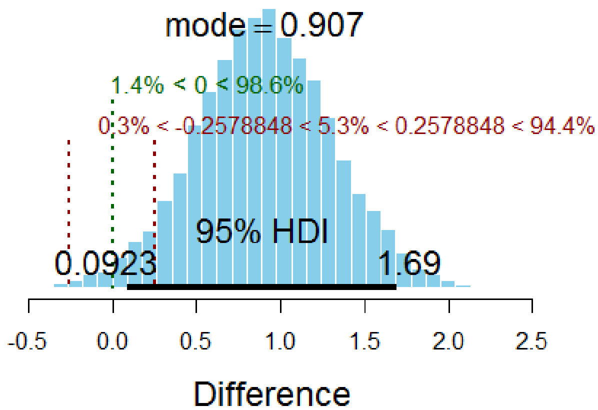

Figure 59. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of taking full advantage of the CP in a timely and efficient manner.

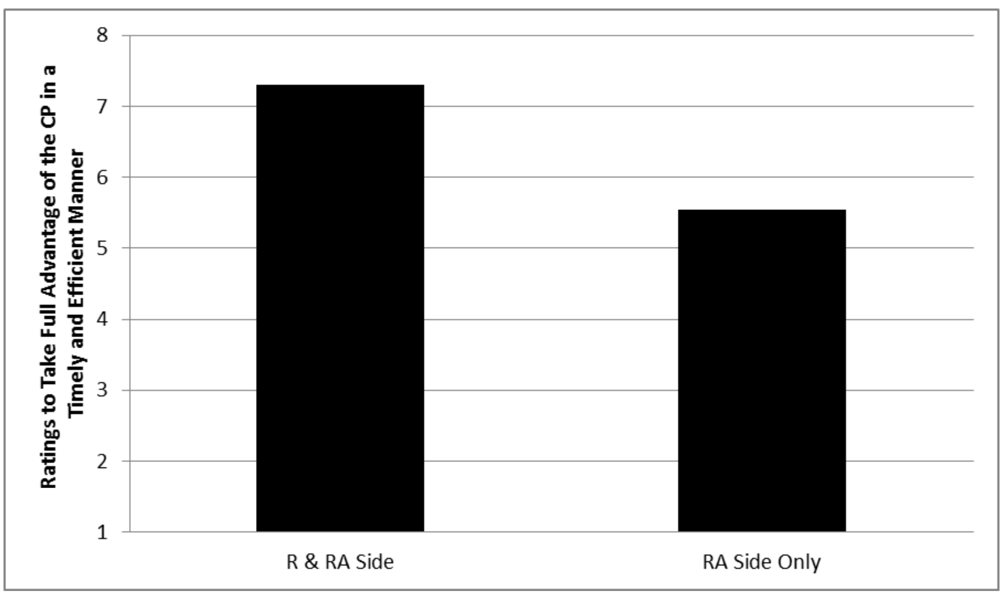

Figure 60. Graph of the mean estimates from the CP Location analysis of the ratings of taking full advantage of the CP in a timely and efficient manner.

For overall CP use, an HDI ranging from 0.604 to 1.855 indicated that CP Location influenced observer ratings (see Figure 61 for ROPE and comparison to zero). Figure 62 illustrates that ratings were higher when the CP was on the R- and RA-Side.

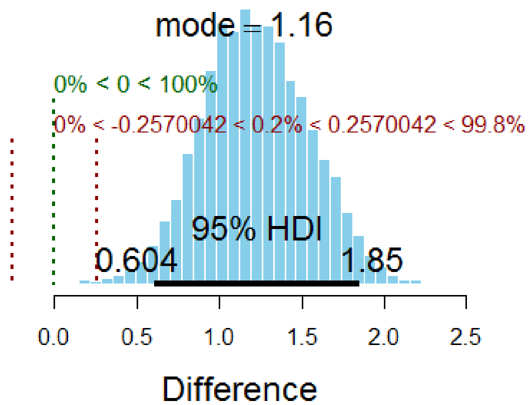

Figure 61. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of overall CP use.

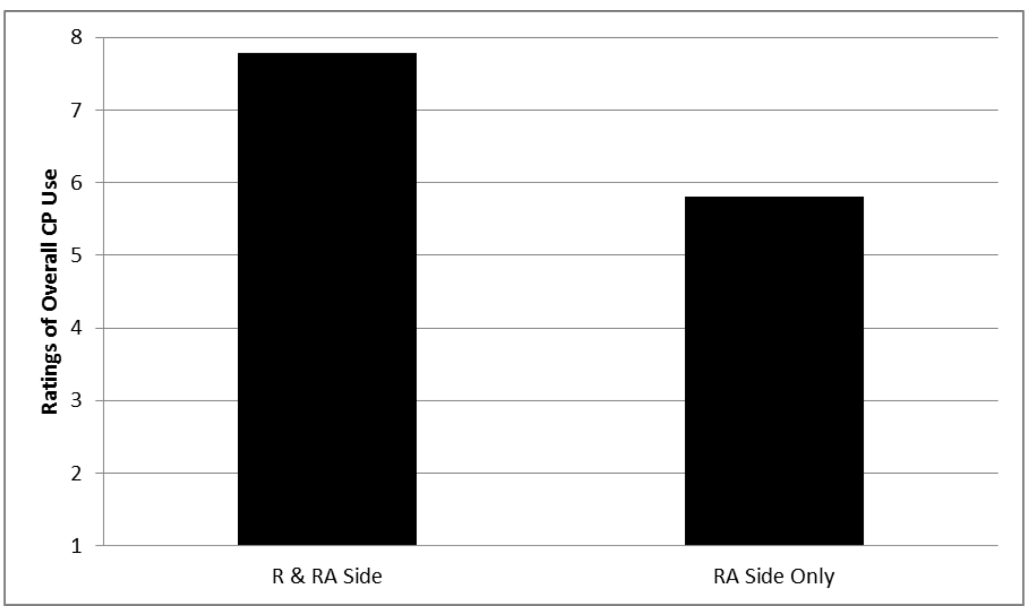

Figure 62. Graph of the mean estimates from the CP Location analysis of the ratings of overall CP use.

3.3.2.2 Radar Surveillance Noise by CP Location Analysis

With the over-the-shoulder rating form, observers often gave controllers consistently high ratings on a number of different outcomes. There was no effect of the manipulations for these outcomes, but a large percentage of credible responses were within the ROPE. To simplify this reporting, we summarize these results in Table 9 with a list of questions and findings per condition.

Table 9. Summary of Over-the-Shoulder Ratings with a Significant Percentage of Credible Values within the ROPE for Radar Surveillance Noise by CP Location Analysis

	% Credible Values within ROPE		
Over-The-Shoulder Rating Item	**CP Location**	**Radar Surveillance Noise**	**CP Location by Radar Surveillance Noise**
3. Maintaining Safe and Efficient Traffic Flow: Using Control Instructions Effectively	99.9%	99.9%	96.8%
7. Maintaining Attention and Situation Awareness: Detecting Pilot Deviations from Control Instructions	96.8%	96.9%	75.1%
14. Providing Control Information: Providing Essential ATC Information	99.6%	99.7%	97.6%
15. Providing Control Information: Providing Additional ATC Information	98.9%	98.5%	94.6%
16. Providing Control Information: Overall Providing Control Information	99.0%	99.3%	95.7%
17. Technical Knowledge: Showing Knowledge of LOAs and SOPs	94.9%	96.9%	82.0%
18. Technical Knowledge: Showing Knowledge of Aircraft Capabilities and Limitations	100%	100%	100%
19. Technical Knowledge: Overall Technical Knowledge	100%	100%	96.7%
20. Voice Communications: Using Proper Phraseology	100%	100%	100%
21. Voice Communications: Communicating Clearly and Efficiently	100%	100%	100%
22. Voice Communications: Listening for Pilot Read backs and Requests	100%	100%	100%
23. Voice Communications: Overall Voice Communication	100%	100%	100%

With respect to using the CP in appropriate situations, an HDI ranging from 0.663 to 1.960 indicated that CP Location impacted observer ratings (see Figure 63 for ROPE and comparison to zero). Figure 64 illustrates that ratings were higher when the CP was on the R- and RA-Side.

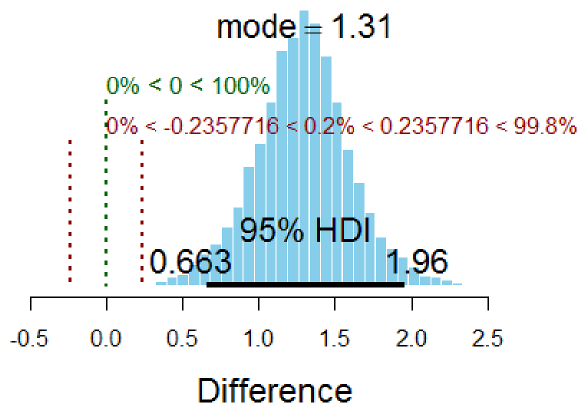

Figure 63. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of using the CP in appropriate situations.

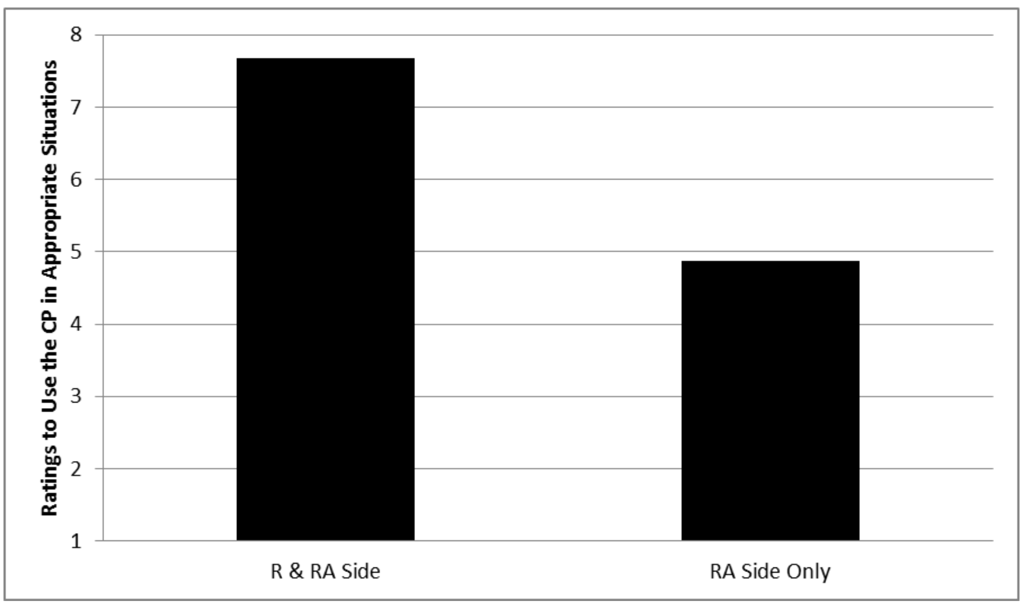

Figure 64. Graph of the mean estimates from the CP Location analysis of ratings of using the CP in appropriate situations.

Observers rated whether controllers took full advantage of the CP in a timely and efficient manner. The analysis of these ratings found an HDI that ranged from 1.056 to 2.336 (see Figure 65 for ROPE and comparison to zero), the manipulation to CP Location influenced observer ratings of overall CP use by controllers. Figure 66 illustrates that observers rated controllers' CP usage higher when it was on the R- and RA-Side than the RA-Side alone.

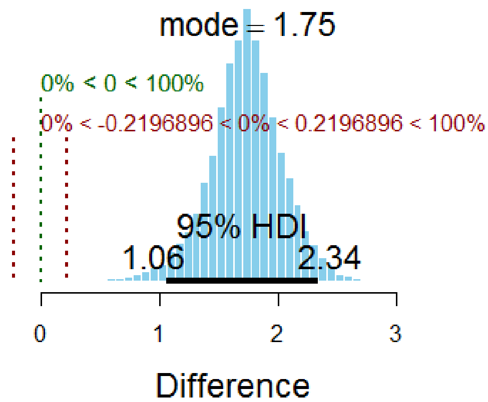

Figure 65. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of taking full advantage of the CP in a timely and efficient manner.

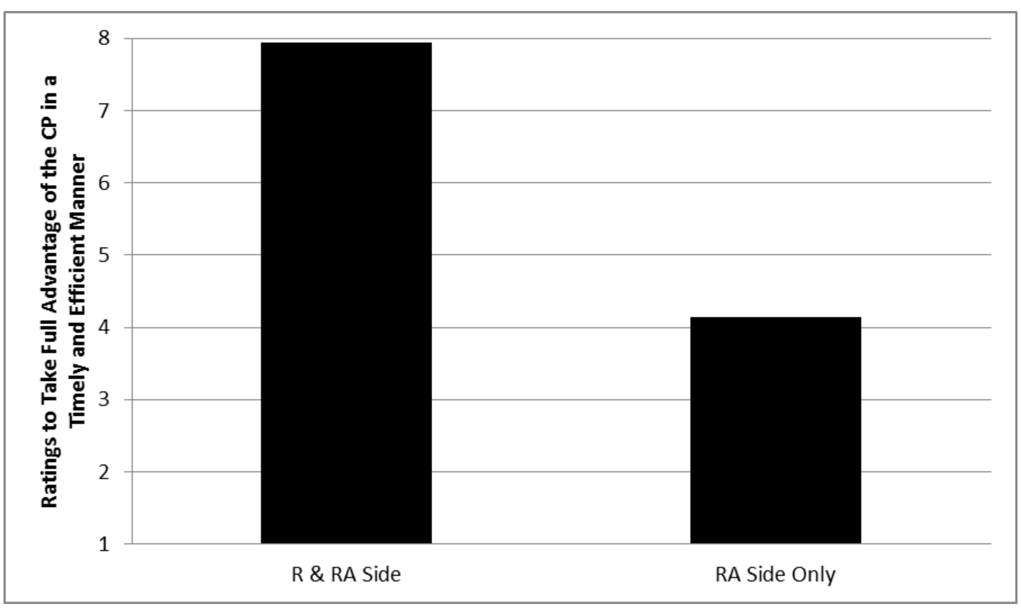

Figure 66. Graph of the mean estimates from the CP Location analysis of the ratings of taking full advantage of the CP in a timely and efficient manner.

With an HDI that ranged from 0.794 to 2.268 (see Figure 67 for ROPE and comparison to zero), the manipulation to CP Location influenced observer ratings of overall CP use by controllers. Figure 68 illustrates that observers rated controllers' CP usage higher when it was on the R- and RA-Side than the RA-Side alone.

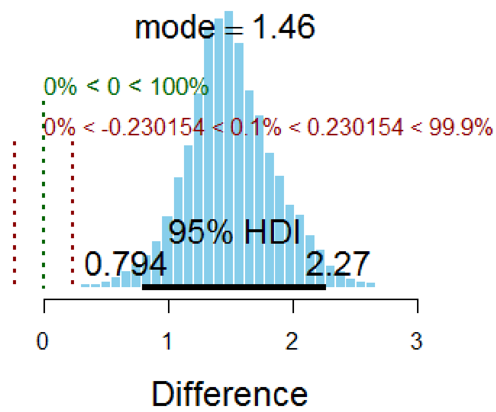

Figure 67. Posterior distribution displaying a 95% HDI, comparison to zero, and ROPE for the CP Location analysis of ratings of overall CP use.

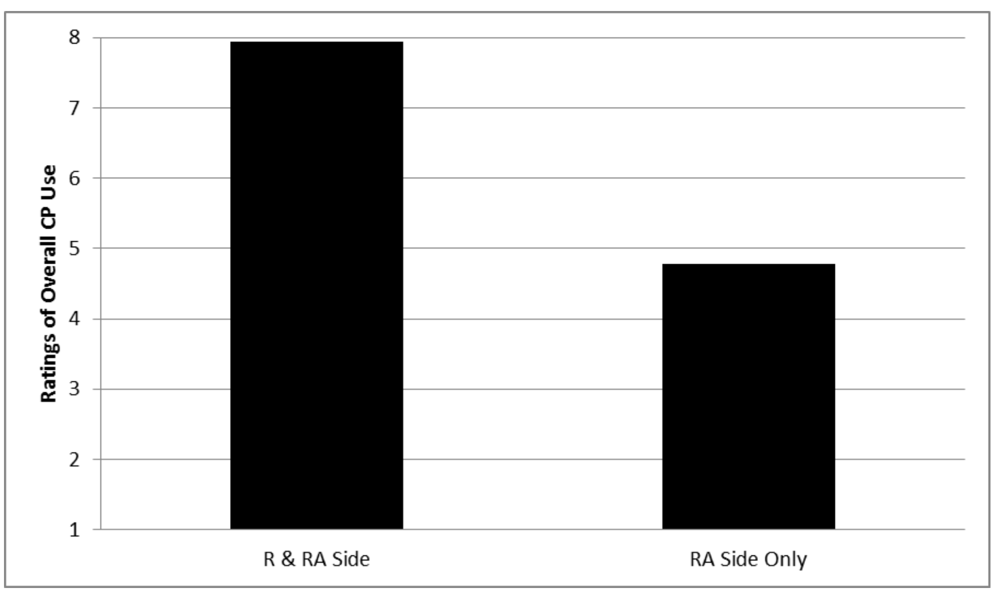

Figure 68. Graph of the mean estimates from the CP Location analysis of the ratings of overall CP use.

3.3.3 Exit Questionnaire Rating Form

At the end of the experimental sessions, we administered an Exit Questionnaire (see Appendix D). We analyzed these responses comparing Group 1 (i.e., Participants 1 to 4) and Group 2 (i.e., Participants 5 to 8) of controllers. When asked to rate the realism of the simulation software, the analysis indicated that controllers in the second group rated the realism higher than participants in the first group, $t(6) = 3.66$, $p = 0.01$. We also observed a significant difference in how groups rated the accuracy of the CP, where ratings of CP accuracy were higher in Group 1 than Group 2, $t(6) = 2.83$, $p = 0.03$. See Appendix K for a full list of means, standard deviations, and test statistics.

3.3.4 Debriefing Comments

This section reviews highlights from the debriefing comments. We did not statistically analyze these data or conduct formal content analysis on it, but we did use it to highlight controller attitudes and comments regarding the primary use of alerts on the CP.

Controllers commented that they predominately used the R-Side to observe and control traffic. One controller simply stated, "Don't use the D-Side [the RA-Side] as much." Another controller said, "I always missed because R-Side didn't alert them," indicating that the controller did not see yellow alerts that the R-Side did not display.

Controllers said that the CP was useful when it was presented on the R-Side with algorithmic improvements. When asked if there was a benefit to only showing red alerts on the R-Side, two controllers said that they saw a benefit. One said it was "much [better]," and the other said it was "…more than acceptable. A benefit."

Regarding yellow alerts, controller comments illustrated two major points. First, there were too many yellow alerts on the R-Side display in the conditions with legacy algorithms, and two of the controllers said it was a "waste." The second point was that some controllers still wanted to see yellow alerts, which one controller said was "useful." In the debriefing sessions, five controllers across both groups mentioned the potential benefit of presenting yellow alerts outside of the 5 NM threshold, as long as the algorithmic improvement algorithms were in place, but commenters exhibited a range on the exact distance they would prefer:

- "Would want to match Conflict Alert, which is 5.5"
- "Six miles is enough to grab your attention"
- "I would go to 8 on yellow"

3.4 Safety and Controller Behavior Data

We used the in-house simulation review and replay tool SimViewer, and/or VLC Media Player, to review selected portions of relevant experiment runs. The information available for review includes video displays of the R-Side and RA-Side positions, and an over-the-shoulder camera. The videos contain audio of all controller and pilot communication, as well as communications made by the controller to the observer even when the controllers did not key the microphone. When replaying selected runs, we cross-referenced what was happening in the replays with data from the experiment that were available via SimViewer and/or separate reports produced during the data reduction process. This reference data included the controller-entered commands, simulation pilot commands, and aircraft track updates.

3.4.1 Loss of Separation

Only one loss of separation occurred during the experiment; it happened in a run when the Conflict Probe was only on the RA-position. The sequence of events, as observed in an audio and video replay of the run, was as follows; all times are in simulation time from the beginning of the run, in minutes and seconds:

- 29:50 – Participant descended DAL692 at FL310 to FL240 through N205SP at FL300, which had a muted yellow CP Alert on the RA-Side; and ASQ6084 at FL290, which had a muted red CP Alert. The alerts turned to red alerts when FL240 was entered.
- 30:14 – Conflict Alert (CA) activated between DAL692 and N205SP as well as between DAL692 and ASQ6084.
- 30:21 – Participant expedited DAL692 through FL290 to avoid conflict with N205SP.
- 30:29 – Participant turned DAL692 to a 300 heading to avoid ASQ6084; at the time, the separation between the two aircraft was only 500 feet vertically, and the CA view showed a 4.29 NM lateral separation.
- 30:42 – Participant says, "Didn't see that."

3.4.2 Deviations

Our data reduction process enabled us to assess the prevalence of two types of operational deviations. The first was frequency deviations, and the second was instances of failure to meet the altitudes required by SOPs.

3.4.2.1 Frequency Deviations

We assessed the percentage of flights for which participants transferred communications to the next sector using the correct frequency and in a timely fashion; that is, the verbal transmission was completed prior to boundary crossing. In general, we determined correct and timely communications transfer by automatically comparing corresponding records for the same flight in the simulation pilot command data and the boundary crossing times, both of which are recorded by our simulation environment. If the pilot entered a communications transfer to the frequency of the next sector, and did this before the boundary crossing, then it could be assumed that the controller had issued the correct frequency at the appropriate time.

We reviewed a subset of the other cases to determine whether controller error had caused the late/incorrect transfer. If the automated assessment showed that an incorrect frequency was issued, or that no frequency change took place, we reviewed the playback. In some of these cases, the controller had in fact issued the frequency change correctly, but the pilot had either failed to enter it at all or had entered the wrong frequency. Simulation pilot entry of the wrong frequency was relatively rare, because the pilots had macros on their workstations for all the commonly used frequencies in this experiment and did not have to manually type the frequency change command. Any event that we determined to be due to pilot error or anomalous behavior by the simulation environment was removed from analysis. However, in some cases, the incorrect or missed communications transfer was attributable to controller error and would be classed as a deviation in the field. We scored these as unsuccessful transfer of communications (frequency deviation) events.

For the records that remained, where the transfer was to the correct frequency but the pilot entry took place after sector boundary crossing, we considered that some time would naturally elapse between when the controller instructed the pilot to switch to the next frequency until the time the pilot read the clearance back and entered the command into the pilot interface. Therefore, our automated comparison could show a late transfer of communications even though the participant issued the verbal transfer prior to the flight exiting the sector. To determine the amount of time after boundary crossing that could be considered timely, we sorted all flight records on the time lapse between boundary crossing and simulation pilot command entry time. We reviewed command entries occurring between 5.7 and 30 seconds after sector exit by replaying the simulation video/audio recordings to assess where the aircraft was when the clearance was issued. We chose these cutoffs because when sorted by the time differential, 20 consecutive clearances where the data showed less than a 5.7-second time lapse were verified to be timely and 20 consecutive clearances showing as taking place more than 30 seconds after crossing were verified to be late. All clearances outside these bounds that we had not manually reviewed were scored as timely for a lapse of less than 5.7 seconds and scored as late for a lapse of greater than 30 seconds. Because pilot entries happening between 5.7 and 30 seconds after crossing were ambiguous as to timeliness, we reviewed the replays of all of these events and manually assigned each as either late or timely. We also reviewed several cases with extreme values—either many minutes before or after boundary crossing—and scored them manually.

For the six conditions analyzed for the bulk of the data analysis effort, and the core period of each run falling between the 13-minute and 49-minute marks of the baseline scenario, we first excluded flights that experimenters or the data reduction process had identified as behaving anomalously. We had 2,272 flights remaining that exited the experimental sector and required a transfer of communications. Of these, the controllers transferred 1,835 (80.8%) to the correct frequency in a timely manner. The other 437 events comprised 415 cases of late switches to the right frequency, 2 timely switches to the wrong frequency, 4 late switches to the wrong frequency, and 16 cases of not switching the aircraft at all. Because the vast majority of the unsuccessful transfers were late switches to the right frequency, we classified all events as simply unsuccessful or successful. We also included in the count of unsuccessful frequency transfers those cases where the participant issued the wrong frequency, issued the right frequency, but failed to correct a bad pilot read back, or never issued a frequency change at all for an exiting flight. Figure 69 shows the percentage of flights that were successfully transferred as a function of CP Location: 82.75% when CP was only on the RA-Side and 78.96% when CP was on both the R- and RA- Side.

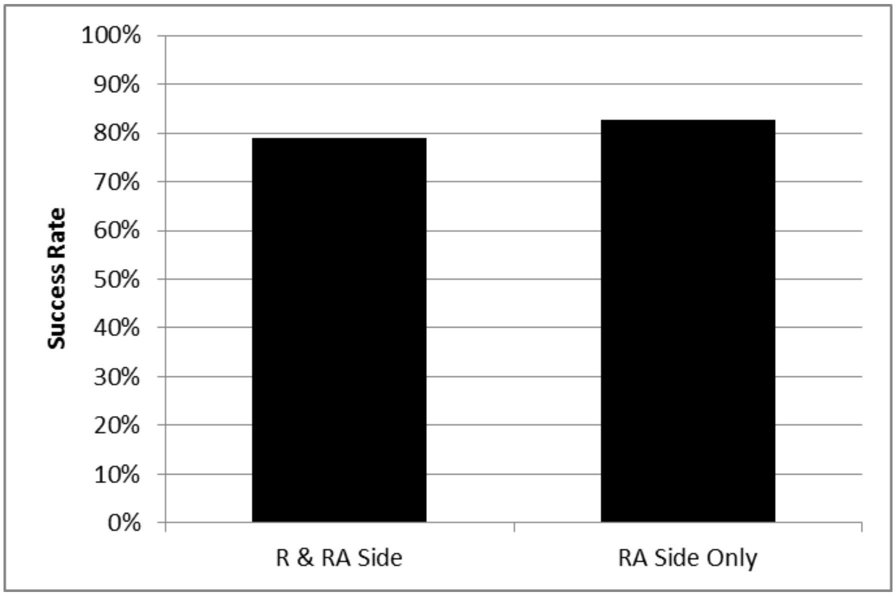

Figure 69. Success rate at frequency switches by location.

A Chi-square analysis of the number of flights successfully transferred on time revealed a significantly higher number of successful transfers when the CP was only on the RA-Side, $\chi^2(1, N = 2272) = 5.26, p < .022$. We conducted Chi-square analyses for this report using a utility by Preacher (2001).

3.4.2.2 Standard Operating Procedures

We assessed the percentage of the flights landing at the three airports whose arrivals were subject to altitude restrictions, CMH, CVG, and CAK, that participants successfully descended to the required altitude before they crossed the sector boundary. We did so by replaying the video of the sector exits of every flight subject to one of the SOPs. If the controller never assigned the aircraft the appropriate altitude, or if the assignment was made but the aircraft crossed the boundary not yet level at that altitude (for CMH and CVG arrivals; as the CAK restriction only required that the aircraft be descending to the restriction altitude when it crossed the boundary), we conducted further examination to determine the cause. If a simulation environment problem had not affected the aircraft and the simulation pilot had made no error, then the event was flagged for further investigation. For these cases, we reviewed the audio to determine if verbal “coordination” took place—that is, the participant stated aloud that they were coordinating the crossing with the receiving sector, prior to the aircraft entering the sector. If so, we did not consider it a violation of the SOP. We considered all remaining incidents to be violations.

For the six conditions analyzed for the bulk of the data analysis effort and the core period of each run falling between the 13-minute and 49-minute mark of the baseline scenario, after excluding flights that had behaved anomalously, we had 268 flights that were subject to one of the restrictions. Of these flights, the controllers descended 241 (89.9%) to the correct altitude in a timely manner. The 27 unsuccessful SOP compliance events included approximately 20 cases that the participant did not issue a descent clearance at all to the exiting flight; in the other situations, they did not descend them to a low enough altitude—level at FL240 (understood in en route operations to mean at or below FL240) for CMH and CVG arrivals and at or descending to FL270 or lower for CAK. As with the frequency deviations, we classified these events in a binary manner as unsuccessful or successful. Figure 70 shows the percentage of flights that controllers successfully descended to meet the restriction as a function of CP Location: 92.42% when CP was only on the RA-Side and 87.50% when it was on both the R- and RA-Side.

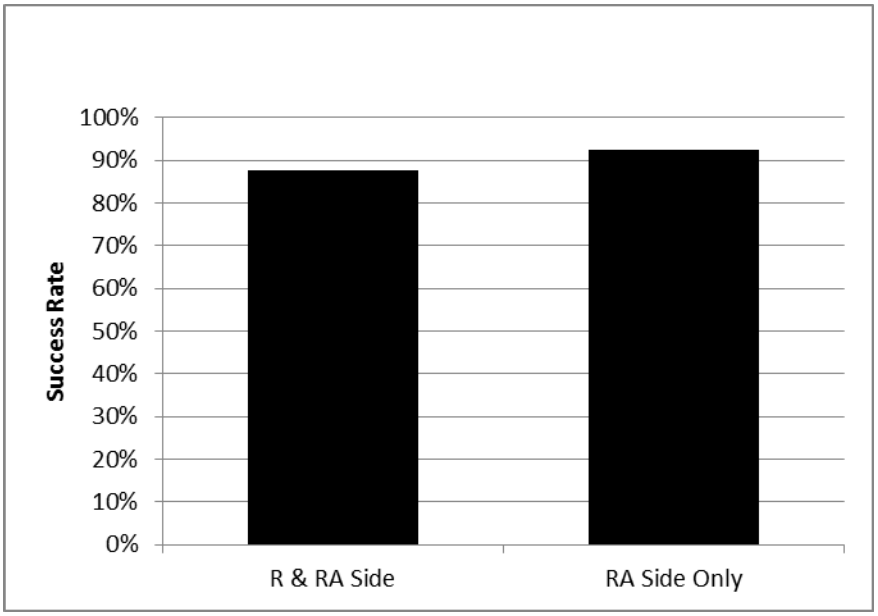

Figure 70. Success rate at meeting altitude restrictions by location.

A Chi-square analysis of the number of flights successfully descended in a timely fashion revealed no significant difference based on CP Location, $\chi^2(1, N = 268) = 1.79, p < .19$.

In addition to the 27 clear cases of failure to meet the restriction, one participant habitually descended a large number of flights—not only those landing at the three airports with SOPs—to FL220, below the floor of ZID85, and his or her comments made it clear that the strategy was a means of shedding workload. We reclassified (as restriction violations) the 10 instances of this behavior across that participant's six analyzed runs for the flights subject to the restrictions, on the basis that the participant may have descended the aircraft to reduce his or her effort in the experiment rather than to comply with the restriction. With this reclassification, the pattern of results was the same across participants, with an 88.6% success rate when CP was only on the RA-Side and 84.6% when CP was on both the R- and RA-Side. This difference was also not significant.

3.4.3 Altitude Clearance Types

We also investigated controller behaviors that we believed would be especially illuminative regarding controller use of CP, and benefits, or lack thereof, including CP functionality on the R-Side. We focused on the following three types of behavior:

1. Expedited altitude clearances
2. Amended altitude clearances (given and then corrected because of a conflict)
3. Intended altitude clearances (changed prior to the clearance being given)

The altitude clearance behaviors covered here occurred a limited number of times over the course of the experiment, and the only variable of interest that we believed would significantly affect these behaviors was Probe location—which was run as planned in all runs and was not affected by the simulation errors described in Section 2.10. Therefore, all the analyses in this section used the data from all 12 experimental runs per participant, and they all included non-core times of each run. The exception is that simulation pilot and controller command data were corrupted for Run 4 by Participant 1, which did not allow it to be included in the expedited or modified clearances analyses in the following subsections. This run was a condition excluded from most of the analyses reported in other sections (A-R5) and so the corrupted data does not affect those analyses.

3.4.3.1 Expedited Clearances

The order to "Expedite" the execution of a clearance is "Used by ATC when prompt compliance is required to avoid the development of an imminent situation" (FAA, 2014). Most controllers use it only sparingly, often to resolve the development of a potential conflict. Therefore, we conducted an in-depth examination of cases where the data showed a participant issuing this type of clearance. We identified these clearances by filtering the simulation pilot commands. We investigated every Expedite command, which were all for altitude clearances. We replayed the run in which the command took place to confirm whether the Expedite was issued by the controller as opposed to the pilot independently deciding to expedite the transition. Pilots entered Expedites independently when they were late entering the command due to workload, or by mistake, or for other unknown reasons. After removing the commands that were not controller-initiated, 22 occurrences remained where the controller actually used the word "expedite" in the clearance, or used phraseology that clearly indicated the intent to have the pilot quickly change altitude; for example, "No delay down," "Increase your rate of descent," and "Good rate on down to 28 please."

We evaluated these 22 occurrences to assess whether controller actions were proactive or reactive. This call was subjective but was based on the operational knowledge of one of the authors—a controller with 27 years of experience at a Level 12 En Route FAA ARTCC and three years' military ATC experience. Factors considered in determining appropriateness of the Expedite included whether the assigned altitude had been available sooner, whether the clearance to expedite was included in the initial clearance or given at a later time, and CA activation timing with respect to the clearance. CA activation before the clearance suggests that the Expedite was in response to the CA, but if it took place after the clearance, it generally means that the clearance was appropriate but the participant gave it at a time where a normal or slow descent may have developed into an unsafe condition. If no CA occurred at all, it suggests that the clearance resolved the conflict in a timely fashion—or that it was unnecessary, but this factor was considered in making the proactive/reactive determination. Each incident had different factors at play and we needed to analyze them on a case-by-case basis, but generally if the controller issued the clearance reasonably soon after the requested altitude became available, and if the Expedite was part of the initial clearance, it was determined to be proactive rather than reactive behavior.

We tallied the number of proactive and reactive Expedites according to whether they occurred where a CP Alert was provided on the R-Side. In the RA-Side Only conditions, this, of course, was never the case. In the R- and RA-Side conditions, sometimes a CP Alert was present on the R-Side, but sometimes it was not, either because the alert was yellow and the condition was not one that showed yellow and muted alerts on the R-Side, or the time to display the alert based on the likelihood function had not yet been reached. Under either of these circumstances, a Probe alert would have been showing on the RA-Side.

For two of the incidents, there was no alert on either the R- or RA-Side. Replays of these events revealed that the aircraft given the Expedite was climbing, and if it did not climb as fast as possible, it was reasonable to conclude that a conflict might occur with crossing level traffic. Therefore the participant was deemed as having been cautious but not unreasonably so, and as having appropriately detected a potential conflict without an R-Side CP Alert—proactive behavior.

Figure 71 shows the counts of Expedited clearances classed as Proactive and Reactive as a function of whether an alert for the particular aircraft expedited was present on the R-Side.

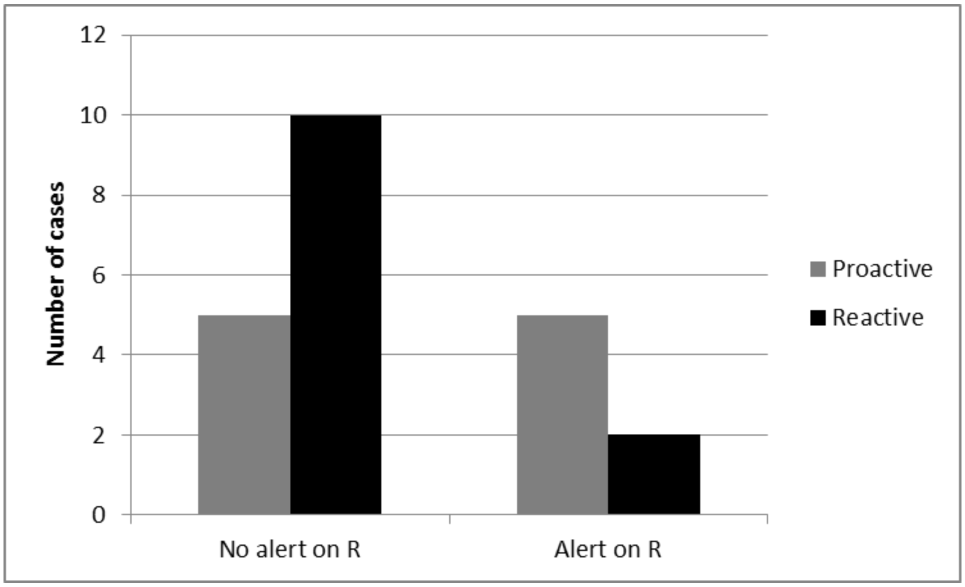

Figure 71. Proactive/reactive classification of Expedited clearances according to presence or absence of R-Side CP Alert.

Expedite was used in a proactive manner 5 of the 15 times when no alert was displayed on the R-Side, and 5 of the 7 times when a CP Alert was present on the R-Side. We conducted a Chi-square analysis on these data and found, $\chi^2(1, N = 22) = 2.79, p < .095$. However, because some of the expected frequencies in the individual cells were less than 5, we applied Yates' correction (Yates, 1934) which yielded, $\chi^2(1, N = 22) = 1.47, p < .23$. There was no statistically significant difference in the proportion of proactive versus reactive Expedite clearances between the R-Side CP present and absent cases.

We also looked separately at the most safety-oriented aspect of these incidents, that is, whether CA activated prior to the Expedite and hence we presume that the clearance was a highly reactive response to the tactical alert. Figure 72 shows these results.

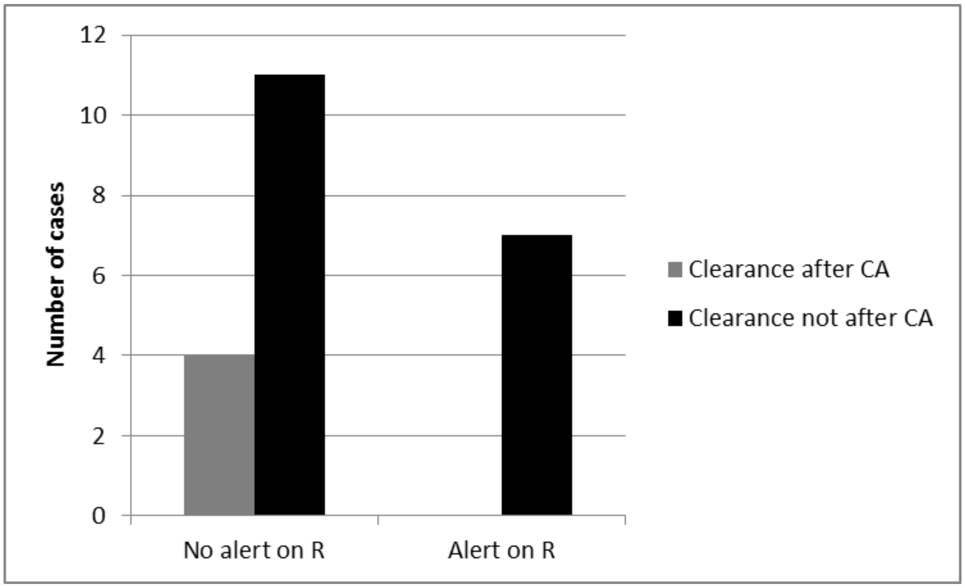

Figure 72. Whether expedited clearance was given only after a CA, according to presence or absence of R-Side CP Alert.

In 4 of the 15 Expedited clearance cases with no CP Alert on the R-Side, the Expedite was given only after the CA appeared. When a CP Alert appeared on the R-Side, in none of the seven cases did a CA develop before the controller issued the clearance. We conducted a χ^2 analysis on these data as well; the result was, $\chi^2(1, N = 22) = 2.28$, $p < .14$. With Yates' correction, $\chi^2(1, N = 22) = 0.85$, $p < .36$. There was no statistically significant difference in the proportion of Expedited clearances taking place only after a CA clearance between the R-Side CP present and absent cases. It is operationally notable, however, that throughout the experiment, no situation for which an R-Side CP was provided developed to the point where a participant deemed it necessary to give an Expedite command in response to a short-term CA.

For an additional look at Expedited clearances' relation to CAs, we classified all cases where a CA occurred into one category regardless of whether it happened before or after the clearance. We based this classification on the logic that if a situation triggering a CA develops, even one that the controller was aware of prior to receiving the tactical alert, the CA still might not have occurred had the controller been more timely with the resolution. Figure 73 illustrates these results.

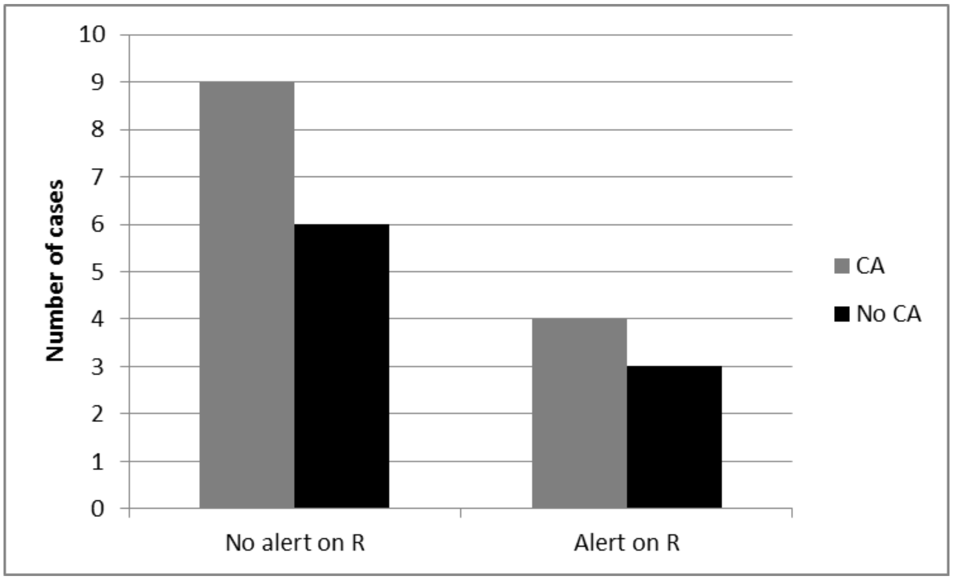

Figure 73. Whether a CA occurred, according to presence or absence of R-Side CP Alert.

With the cases classified in this way, a CA occurred (sooner or later) 6 of the 15 times that there was no CP on the R-Side, and 3 of the 7 times when the CP Alert did display on the R-Side. The differences between the conditions were not statistically significant, $\chi^2(1, N = 22) = 0.016, p < .90$. With Yates' correction, $\chi^2(1, N = 22) = 0.12, p < .74$.

Although there were not enough observations to derive statistically significant conclusions, the analysis of the Expedited clearances provides anecdotal evidence that controllers may have been able to act more proactively when Conflict Probe was present on the R-Side.

Details on the 22 Expedited clearance observations going into the analysis described in this subsection can be found in Appendix L.

3.4.3.2 Clearances Modified after Issued

We examined incidents in which a controller issued an altitude and then later took back all or a portion of the descent or climb clearance(s). As with the Expedited clearances described above, an amended altitude in itself is not an indication of a poor decision or an error by the controller. In fact, amending a clearance, if done in a timely fashion, indicates a correct decision and shows that the controller corrected a suboptimal decision before it became an issue.

As with the Expedites, we initially identified these situations by examining the simulation pilot commands. We used the replay tools to investigate every altitude command entered that was later changed to take back all or a portion of the climb or descent for the same aircraft, to determine why the altitude amendment was given. In several instances, the commands entered were simulation pilot input errors, commands entered by the simulation pilot on the wrong aircraft, or amendments given to comply with SOP requirements. We excluded these instances from our comparisons because we were only interested in cases where the altitude amendments were due to aircraft conflicts. We excluded times when the amended altitude was one which continued the aircraft's climb above or descent below the conflict aircraft. In these situations it was less certain that the second clearance was prompted or influenced by the CP Alert. It may have merely been the way in which the controller had initially planned to transition the aircraft to its target altitude.

We reviewed 27 cases where assigned altitudes were amended taking back part or all of the original clearance. Reviews showed that three were the result of simulation pilot or controller errors due to miscommunication, and two were changed to comply with the SOP for aircraft landing CAK. Of the 22 remaining cases, there was always at least a yellow alert on the RA position. Thirteen occurred in runs with the CP on the R-Side and nine occurred in RA-Side Only conditions.

However, of the 13 occurrences in the R-Side condition, only in 10 of them was an alert present on the R-Side for the aircraft. Therefore, we classified the other three occurrences as RA-Side Only, which was the situation for the aircraft, resulting in 10 cases with an R-Side CP Alert on the aircraft and 12 without.

The detailed list of all the evaluated occurrences appears in Appendix L. To determine whether having a CP Alert on the R-Side facilitated the correction of suboptimal altitude clearances, we calculated the time elapsed between the first and second altitude clearance. For this analysis, if the aircraft was issued the first clearance by the "previous sector" (in other words, automatically executed by the simulator prior to entry into ZID85 as part of the aircraft's climb or descent profile), then the time that the aircraft initially checked on frequency was used as the time of the initial clearance. During the scenarios, there was no procedure established to communicate with the previous sector so this was the first opportunity the R-controller had to take a control action.

Figure 74 shows the average number of seconds elapsed between the two altitude clearances as a function of whether a CP Alert was present on the R-Side for the aircraft to which the controller gave the clearance.

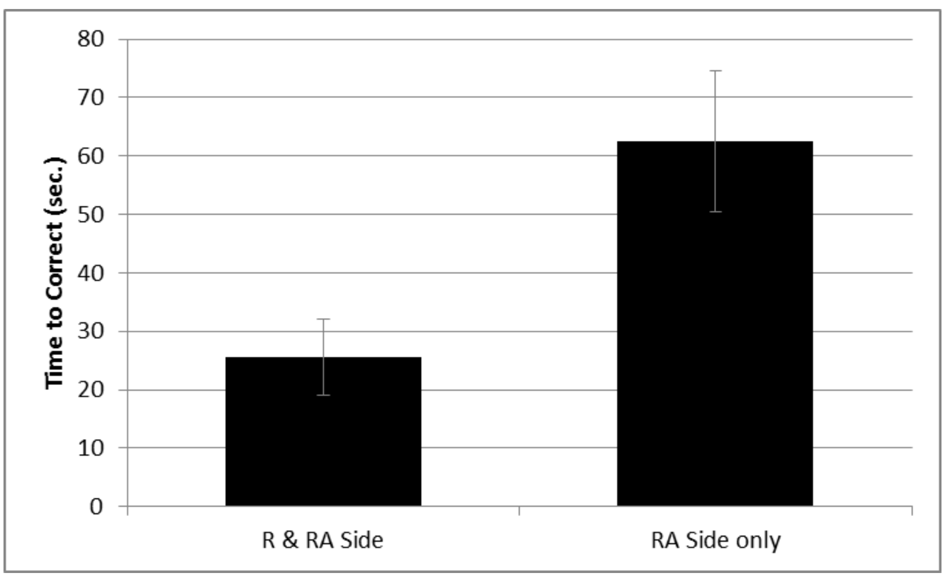

Figure 74. Time elapsed between initial and corrected clearance as a function of CP Location.

In the cases where the amended flight did not have an R-Side CP Alert, it took participants an average of 62.58 seconds to amend the clearance (SD = 41.89). When the CP Alert was available on the R-Side for the flight, the mean time to amend was 25.60 seconds (SD = 20.42), a difference of 36.98 seconds or 144%. This difference was statistically significant according to a within-subjects t-test, $t(19) = 2.50, p < .03$.

3.4.3.3 Clearances Modified before Issued

The final type of altitude clearances that we examined were cases where a controller entered an altitude message into the R-Side workstation but then verbally issued an altitude different from what he or she had just input. These represent situations where the controller used the CP function on the R-Side to modify their initial plan of action before issuing clearances—essentially using the CP information to trial plan future actions. The difference between these situations and the ones described in the previous section is that only one clearance was given to the aircraft; the original clearance considered by the controller was entered into the R-Side but not issued to the pilot.

To discover these entries, we examined sequential controller messages entered into the R-Side for the same aircraft, in which the second command was a modification of the first command. If the order of the altitude entries was an Interim Altitude followed by an Assigned Altitude, this was not included as that is a normal sequence of entries that a controller would make, as for a step-down descent.

The 12 cases we included in this analysis had three things in common:

1. Two different altitudes were entered with only one—or none—being transmitted.
2. Entering the altitude generated a CP Alert in the FDB portal after entering the first altitude (11 red, 1 yellow).
3. None received a CA activation as a result of the clearance.

We computed the time between the first and second message inputs, showing a mean time of 6.00 seconds (SD = 1.88) between the entries (Table L3; see Appendix L).

This analysis did not lend itself to statistical or descriptive comparison between different conditions, because events of this type only took place when controllers had R-Side CP capabilities. It was therefore not possible to assess the difference between presence and absence of R-Side Probe on these events, and there was no operational or theoretical reason to believe that our other IVs would have an effect on Probe-Assisted Decision Making.

One example of how a situation like this played out was when an aircraft at FL330 had a red CP Alert for another aircraft at FL330. The participant ran the vector lines out to 8 minutes, saw traffic at FL350, and said (as though speaking to the aircraft at FL330), "I see thirty four wrong in your future." The participant entered "FL340" into the R-Side workstation, which resulted in a red CP Alert for the aircraft. The controller investigated the alert by displaying the alert trajectory line for the aircraft, laughed, and said, "Oh, that's beautiful," because there would have been two alerted aircraft at FL340. The controller chose to keep the aircraft at FL330 and vectored it 10 degrees right instead of changing its altitude. We also observed several instances of a similar type of situation where a participant entered an altitude into the R-Side workstation and did *not* receive a CP Alert, issuing the altitude to the pilot only after seeing that the intended altitude did not cause a CP Alert.

The descriptive/anecdotal results on this incident type provide anecdotal evidence of another type of behavior fostered by the R-Side CP functions that enabled participants to more proactively make appropriate decisions to resolve aircraft conflicts without creating new ones.

3.5 Objective Workload

3.5.1 PTT

The simulation environment records all presses of the PTT switch, which we used as a measure of controllers' communications workload. We excluded the two longest transmissions (25 seconds or longer) from all analysis, both of which we reviewed and determined were "open mic" situations, and all transmissions less than 0.9 seconds in duration, based on our reviews of 192 transmissions from the experiment as short as .066 seconds. We determined that transmissions shorter than 0.9 seconds never contained meaningful operational ATC information. This determination of a 0.9-second cutoff corroborates the findings of prior reviews for other research conducted by our lab.

We analyzed the use of the PTT switch to obtain an objective measure of communications workload. We used the transmissions occurring during the 35-minute core period of the baseline scenario that was common to all runs.

We assessed the effects of Location, Noise, and Algorithm, and key interactions between them, on the length of time spent making voice transmissions and on the total number of transmissions. We conducted two separate multiple regressions, on transmission time and number of transmissions, for each analysis question—Location main effect, Location and Algorithm and their interaction, and Location and Noise and their interaction. We used the percent of each 2-minute interval, and the number of transmissions during each 2-minute interval, as individual data points. When a transmission straddled multiple 2-minute intervals, we prorated its count among the intervals in which it appeared. For example, a transmission beginning at run time 15:45 and ending at 16:05 included 15 seconds in the 14-16 minute interval and 5 seconds in the 16-18 minute interval and was counted as 0.75 transmissions and 0.25 transmissions in the two respective intervals.

We used the maximum instantaneous aircraft count for the two minutes in the sector being worked, and the number of minutes into the run at which the interval began (time-on-task) as covariates. These variables, along with the independent variable(s) of interest for the given analysis, were entered into the regression equation as fixed effects, and participant as a random effect, to create a within-subjects model. Preliminary comparison of this model with a multilevel within-subjects model using unique run nested within participant, showed that either the non-multilevel model was significantly more reliable, or that the reliability of the two models did not differ significantly and the patterns of significance were the same with either model. Therefore, for each analysis reported in this subsection, we report the non-multilevel results.

We first analyzed the communications workload measures as a function of Location, using the six conditions for each participant where the parameters of the conditions run were approximately as planned. Figure 75 shows the difference between the two Locations: 23.40% (SD = 8.81%), or 28.08 seconds of each 2-minute interval, transmitting on frequency when CP was only on the RA-Side and 24.06% (SD = 8.54%), or 28.87 seconds of each 2-minute interval, when it was on both the R- and RA-Side.

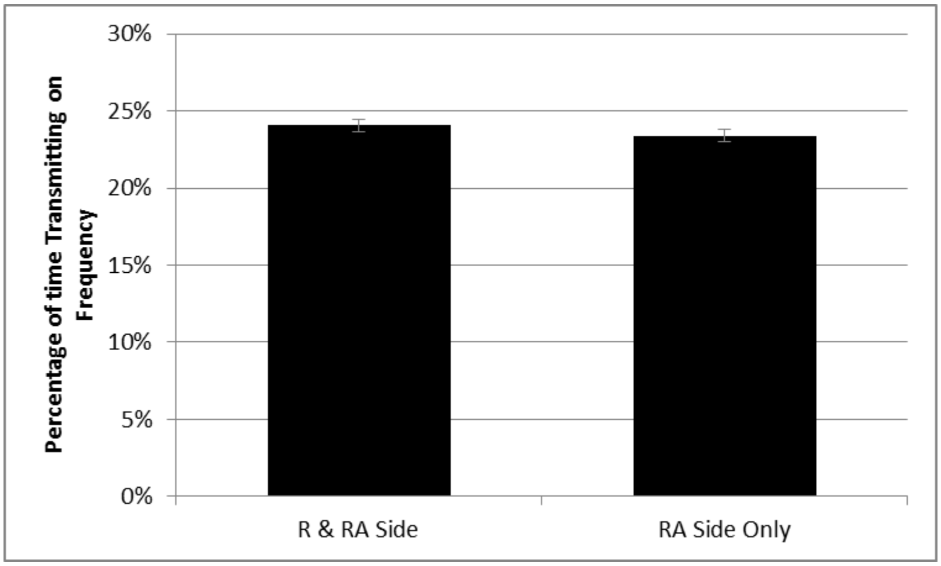

Figure 75. Mean percentage of time on PTT by Location.

The regression model showed a statistically significant effect of Location, $F(1, 805) = 7.33$, $p < .007$.

Similarly, we performed the same regression model on the transmission count data, which appear in Figure 76, converted to the number of transmissions per hour based on the counts per 2-minute interval. The mean hourly transmission rate was 232.05 ($SD = 80.06$) when CP was only on the RA-Side, and 240.06 ($SD = 80.00$) when it was on both the R- and RA-Side.

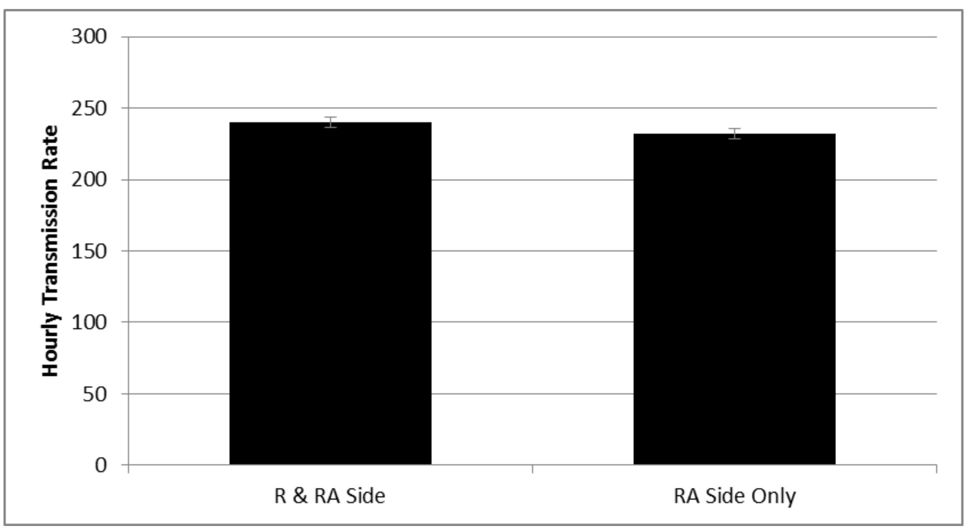

Figure 76. Mean transmissions per hour on PTT by Location.

The regression model showed a statistically significant effect of Location, $F(1, 805.1) = 6.82$, $p < .01$. Communications workload, in total time and total number of transmissions, was higher when CP was available on the R-Side. The pattern was similar to the Location effect reported above, using a smaller subset of the data, and we do not repeat the graphical depiction of the effect here.

When we analyzed the effect of Location and Algorithm on the PTT data, for the duration variable, we again found a significant main effect of Location, $F(1, 532.8) = 9.15$, $p < .003$, with more time on frequency when CP was available on the R-Side. We found no significant main effect of Location, and no interaction.

The transmission count data showed a significant main effect of Location, $F(1, 535.7) = 6.35$, $p < .02$, with more transmissions when CP was available on the R-Side. We found no main effect of Algorithm, and a significant interaction between Location and Algorithm, $F(1, 536.1) = 4.83$, $p < .03$. The combined effect of these two independent variables on the number of transmissions appears in Figure 77.

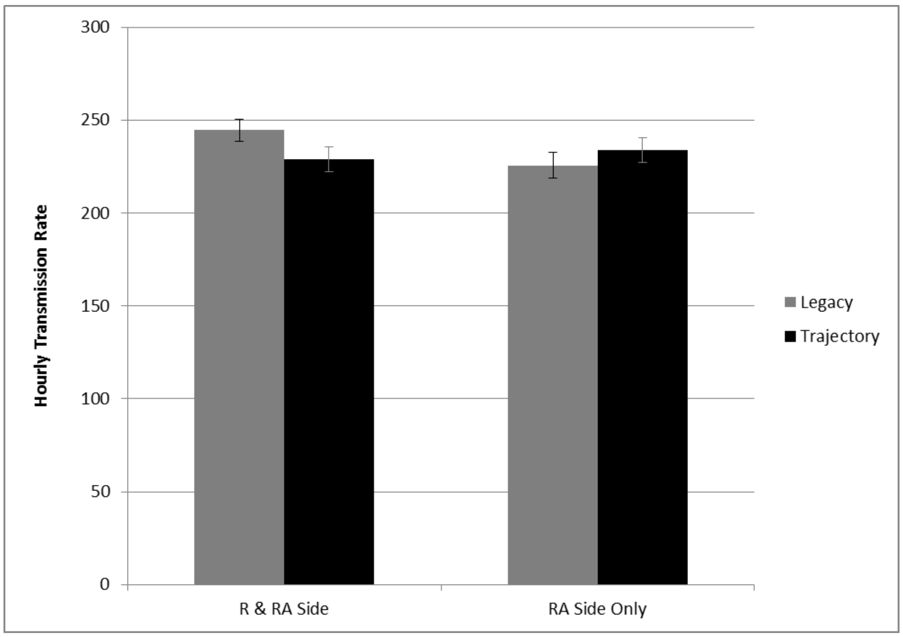

Figure 77. Mean transmissions per hour on PTT by Location and Algorithm.

A post-hoc Tukey test comparing the four means that we analyzed for the interaction revealed that the only two means that differed from each other were the two Legacy conditions. We recorded an average of 19 more transmissions per hour in the R-Side Legacy condition containing the additional yellow and muted alerts ($M = 244.66$, $SD = 71.39$), than during the RA-Side Only Legacy condition ($M = 225.63$, $SD = 82.31$).

We next assessed the Location and Noise effects. For duration, the regression analysis showed a trend toward statistical significance for Location, $F(1, 533.2) = 3.50$, $p < .07$, with more time on frequency when CP was available on the R-Side. We found a significant main effect of Noise, illustrated in Figure 78. Controllers spent more time on frequency when Noise was absent ($M = 23.98\%$, $SD = 8.80\%$) than when it was present ($M = 23.43\%$, $SD = 8.49\%$). The interaction of Noise and Location was not significant for PTT duration.

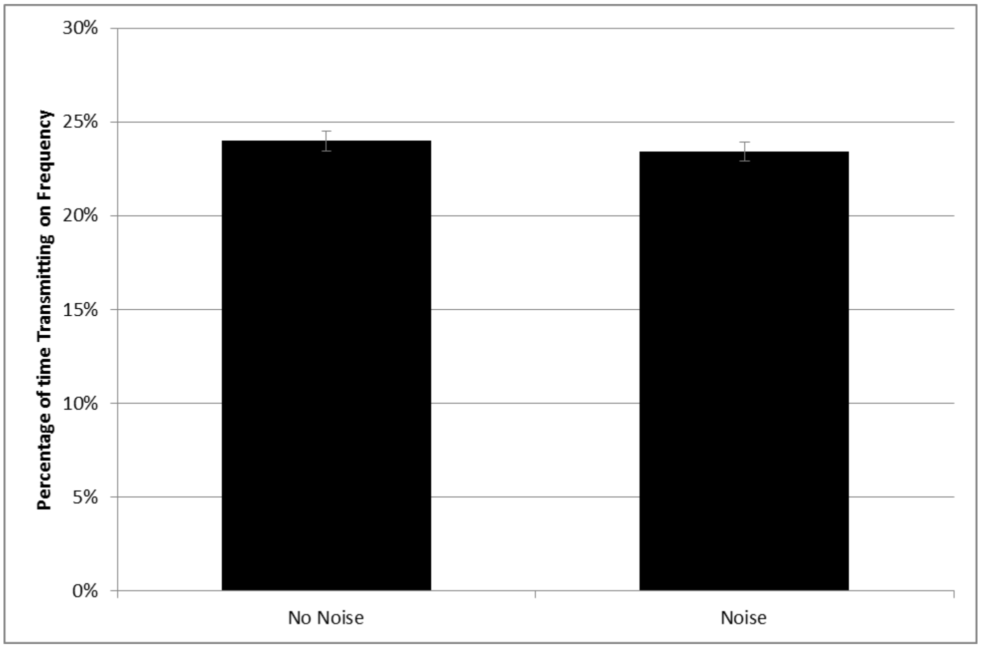

Figure 78. Mean percentage of time on PTT by Noise.

The result of the Noise main effect in the regression model was $F(1, 533.2) = 5.07, p < .03$.

For the Location and Noise effects on transmission count, we did not find a significant effect of Location or of the interaction between Location and Noise. As with duration, we found a significant main effect of Noise (see Figure 79). Controllers made more transmissions per hour when Noise was absent ($M = 241.56$, $SD = 83.36$) than when it was present ($M = 231.47$, $SD = 79.23$).

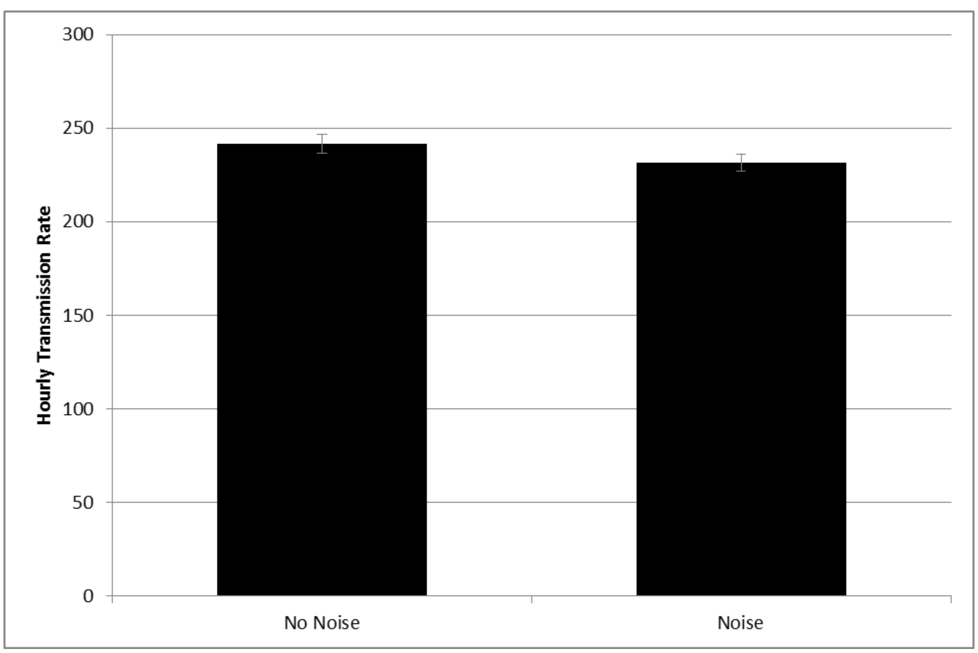

Figure 79. Mean transmissions per hour on PTT by Noise.

The result of the Noise effect in the regression model was $F(1, 536.1) = 7.66, p < .006$. The interaction of Noise and Location was not significant for PTT count. Full details of the PTT analyses are found in Appendix M.

3.5.2 Controller Commands

DESIREE records the time, duration, and exact command text of controller interactions with the workstation. The data collection and reduction process assigns a type to each command, such as Altitude or Speed. As with the WAK analysis and other data types, we analyzed the 2-minute intervals that fell entirely within the core period of the baseline scenario common to all runs. We included in our analysis only commands entered from the R-Side position. A small number of the commands that the simulation recorded (2.7% of commands across all times in all analyzed conditions) were entered on the RA-Side. Observation during the runs and while viewing replays showed that these were usually entered by the over-the-shoulder observers and therefore not measures of participant behavior and workload.

To assess the effects of Location, Noise, and Algorithm, and key interactions between them, we conducted a multiple regression analysis for each of the four main command types—Altitude, Heading, Speed, and Route. The unit of observation was the number of commands given in each 2-minute period. We entered the maximum instantaneous aircraft count in the given period and the number of minutes into the run at which the interval began (time-on-task) as covariates, and participant as a random effect. Preliminary comparison of this model with a multilevel within-subjects model using unique run nested within participant, generally showed either that the non-multilevel model was significantly more reliable, or that the reliability of the two models did not differ significantly and the patterns of significance were the same with either model. Therefore, for each analysis reported in this subsection, we report the non-multilevel results. For ease of interpretation, we convert the command counts to a number of commands per hour of that type.

As is the case in the field, controllers varied widely in how often they entered commands into the console during the experiment. Most notably, two participants never entered Heading commands; we excluded them from the analysis of Heading commands. Also, three participants (the same two plus one other) never entered Speed commands, and we excluded these participants from our Speed command analysis.

We cover each of the four command types in turn. None of the IVs manipulated in the experiment affected the number of Altitude commands entered.

CP Location significantly affected Heading commands. The direction of this effect was that participants entered more Heading commands when CP was available on the R-Side. Appendix N contains the details of this main effect, but the finding should be interpreted in light of the significant interaction between Location and Algorithm. Figure 80 shows this two-way effect.

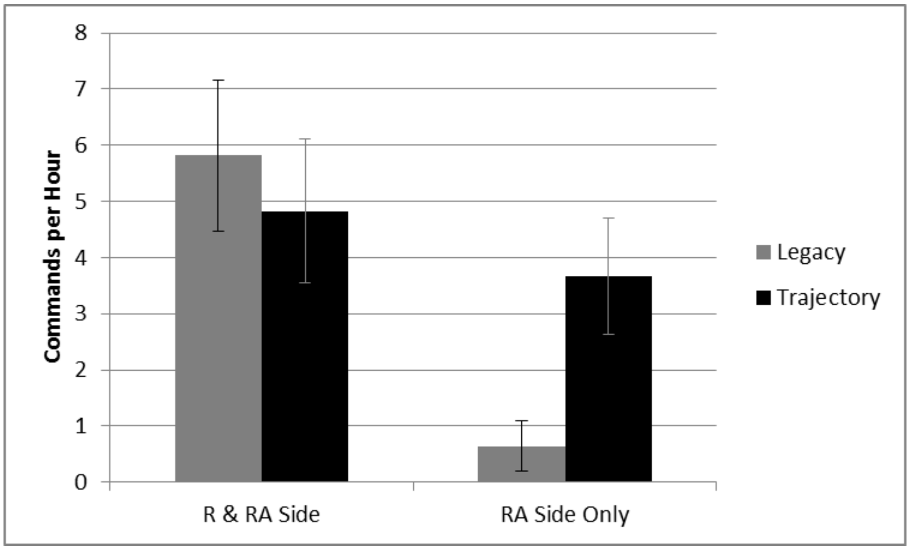

Figure 80. Mean heading commands per hour by Location and Algorithm.

The regression analysis result for the interaction of Location and Algorithm was $F(1, 357.1) = 3.98, p < .047$. A post-hoc Tukey test comparing the four conditions included in the analyses revealed that controllers entered significantly fewer Heading commands in the RA-Only, Legacy condition than in either of the two R- and RA-Side conditions.

Location, Algorithm, and their interaction did not significantly affect the number of Speed commands. We did observe a trend toward a significant effect of Noise (see Figure 81); participants entered more Speed commands per hour in the presence of Noise ($M = 3.06$, $SD = 10.38$) than in the absence of Noise ($M = 1.31$, $SD = 7.01$).

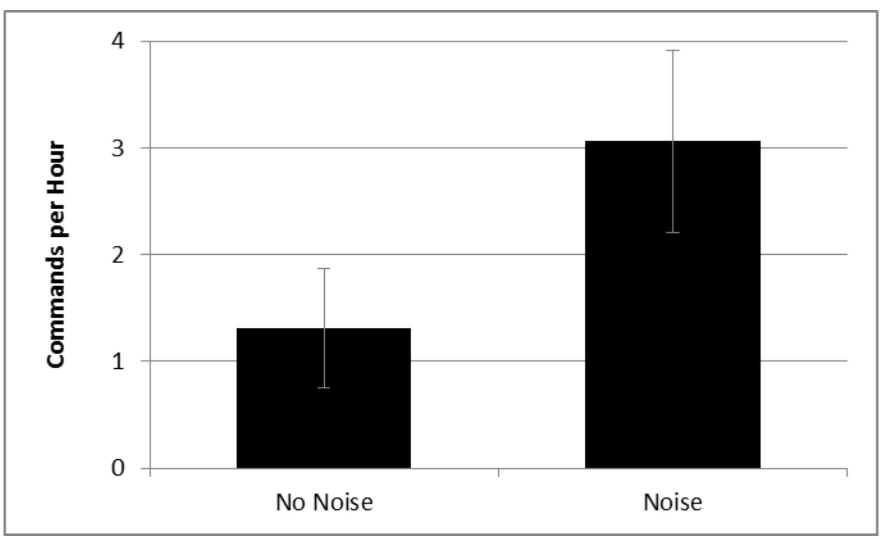

Figure 81. Mean speed commands per hour by Noise.

The regression result showing the trend toward more Speed commands in the Noise condition was $F(1, 297.6) = 3.79, p < .053$.

Location, Algorithm, and their interaction did not significantly affect the number of Route commands, nor did the main effect of Noise. The only significant effect on the number of Route commands was from the interaction of Location and Noise, depicted in Figure 82.

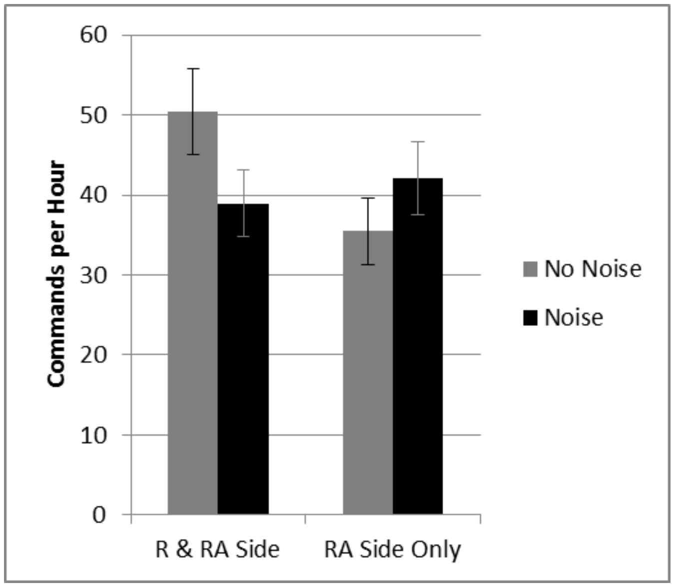

Figure 82. Mean route commands per hour by Location and Noise.

The regression analysis result for the interaction of Location and Noise was $F(1, 474.3) = 6.21$, $p < .014$. A post-hoc Tukey test comparing the four conditions included in the analyses revealed that the only two significantly different conditions were the two No Noise conditions, in which controllers entered more Route commands in the R- and RA-Side/No Noise condition than in the RA-Side Only/No Noise condition.

Appendix N shows the detailed results for the analyses reported in this subsection, and the hourly rates for all individual command types.

3.5.3 Simulation Pilot Commands

As noted in the previous subsection, for certain types of clearances and actions, commands entered into the controller workstation are an incomplete measure of the objective workload due to acting on aircraft under control. Controllers might issue a command but not enter it into the workstation. When this occurs, the simulation pilot will usually enter the command, even if the controller needs to remind them to do so. Therefore, simulation pilot command data is an objective workload measure correlated with, but not identical to, controller commands.

Our data reduction process recorded all commands entered by the pilots and categorized them into categories such as heading and speed commands. Unlike the other data collected continually throughout the experiment, due to the non-normality of this data, we used Bayesian ANOVA to analyze simulation pilot commands.

Our analysis did not identify any significant effects of any of the manipulations. For route clearance commands in both the CP Algorithm by CP Location analysis and the radar surveillance noise by CP Location, we observed little to no variance in the use of these commands, and 100% of the credible values were within the ROPE for all conditions and interactions. See Appendix J for means and standard deviations (*SD*).

3.6 Efficiency Data

3.6.1 Time and Distance in Sector

DESIREE records sector entries and exits, and transfers of track control, and the data reduction process computes the amount of time each flight in the experiment was in the sector that participants were working (and every other sector in the airspace used), the amount of time the flight was under control of the participant, and other metrics of how many flights are contributing to the controller's workload at any time. We analyzed the time aircraft spent in ZID85, and the distance covered as they traversed the sector, as measures of how efficiently participants processed the aircraft through the sector. We excluded from analysis flights exhibiting anomalous behavior and those for which the core time did not include both their ZID85 entry and exit in all runs; that is, they were already in the sector at the beginning of the core time or had not yet exited when it ended. For this experiment, our data reduction produced time and distance metrics according to three definitions: aircraft within the geographical sector limits, aircraft under track control, and aircraft under responsibility, which means the flight is within the sector, under control, or both.

Of the six metrics available, we conducted statistical analysis on the three time metrics. For each of the three definitions of contribution to workload—geographical limits, under control, and under responsibility, the time and distance values were very highly correlated with each other, with r values of greater than .98, so we deemed that analyzing both metrics would contribute no additional information. For our analyses, we used a multilevel approach unique to the time in sector analysis by nesting beacon code within participant. In our scenario design approach, aircraft with the same beacon flew nearly or exactly the same flight path at the same speed and were the same aircraft type from run to run; the same beacon generally differed between runs only in terms of its calling. Therefore, to increase statistical power, we treated each unique beacon code as a factor whose time in the sector and under control would not be expected to differ much from condition to condition except due to participant behavior, such as assigning vectors to the flight to avoid a conflict or giving the flight a direct route for efficiency.

Despite the statistical power of this analysis approach, we found few significant effects. Several of the analyses on Location and Algorithm revealed significant main effects of Location, with slightly shorter times under control and under responsibility when CP was available on the R-Side, but the differences between the mean values were on the order of 1 second, which we consider neither operationally nor theoretically meaningful. Detailed results of the time and distance analysis appear in Appendix O.

3.6.2 Number of Aircraft Handled

As an additional measure of efficiency and behavioral changes due to the various IVs, we analyzed the total number of aircraft processed through the sector in the core time of each run (i.e., the common 35 minutes shared across all runs) and converted it by multiplying by (60/35). We conducted one multiple regression analysis each for Location, Location/Algorithm, and Location/Noise, with the IV(s) of interest as a fixed effect and Participant as a random effect. The main effect of Location was not significant, but the main effect of Algorithm, $F(1, 21) = 6.19$, $p < .022$, indicated more aircraft worked per hour with algorithmic enhancements ($M = 65.57$, $SD = 2.30$) than with Legacy algorithms ($M = 61.71$, $SD = 2.34$). The main effect of Noise was significant, $F(1, 21) = 6.09$, $p < .023$, and indicated more aircraft worked per hour with Noise absent ($M = 65.57$, $SD = 2.30$)—the same condition used in the Algorithm comparison—than present ($M = 63.54$, $SD = 2.03$). We did not consider these differences of about two aircraft per hour, equivalent to a difference of approximately 0.5 aircraft in the 15-minute interval used to gauge sector workload in Traffic Flow Management, to be operationally or theoretically meaningful. Detailed results of this analysis, which we derived from the time and distance data, follows the details of that analysis in Appendix O.

3.7 Conflict Probe Reliability Assessment

To manipulate the reliability of the CP, we performed an engineering experiment similar to those performed by Crowell et al. (2011, 2012) to determine the impact of changes to the minimum adherence bounds and algorithmic changes on reliability. Because we needed a large number of encounters, we used the existing scenarios from Washington (ZDC) and Chicago (ZAU) ARTCCs used by Crowell et al., but we removed surveillance noise. We used the data to determine the lateral and longitudinal adherence bounds that result in the five reliability levels needed in the HITL experiment. Because the reliability levels calculated based on the engineering experiments relied on airspace and traffic samples from ZAU and ZDC, we created an approach to determine the reliability levels that controllers experience during the HITL experiment for two reasons: First, we want to ensure that we achieved the reliability levels that we intended. Second, we want to be able to determine the impact of independent measures other than the reliability manipulation on the reliability values.

The simulation configuration included Indianapolis (ZID) Sectors 95 and 79 combined onto Sector 85. Because of the difference in airspace and traffic between the scenarios that we used to determine the adherence bounds for the five reliability levels, we wanted to determine the CP reliability that controllers experienced during the experiment (see Figure 83). To determine the reliability, we needed to tally the missed alerts, false alerts, correct rejections, and correct detections. One of the challenges was to investigate if a CP Alert would have resulted in a loss of separation given the state of the simulated NAS at the time of the CP Alert. To do that, we used Data Reduction and Analysis Tool (DRAT), provided by the TGF team, to create simulation pilot event files for each of the simulation runs. We reduced the DESIREE data to extract the CP Alerts received during all of the simulation runs. With a script written in R (R Core Team, 2015), we used the CP Alerts data and created a new, truncated pilot event file for every CP Alert that only contained pilot events until the time of the CP Alert. In the fast-time mode of the TGF, we ran the truncated pilot-event files and recorded the TGF data that included aircraft position data. We used DRAT to filter the TGF recordings for the two aircraft involved in the potential conflict and calculated the distance between them.

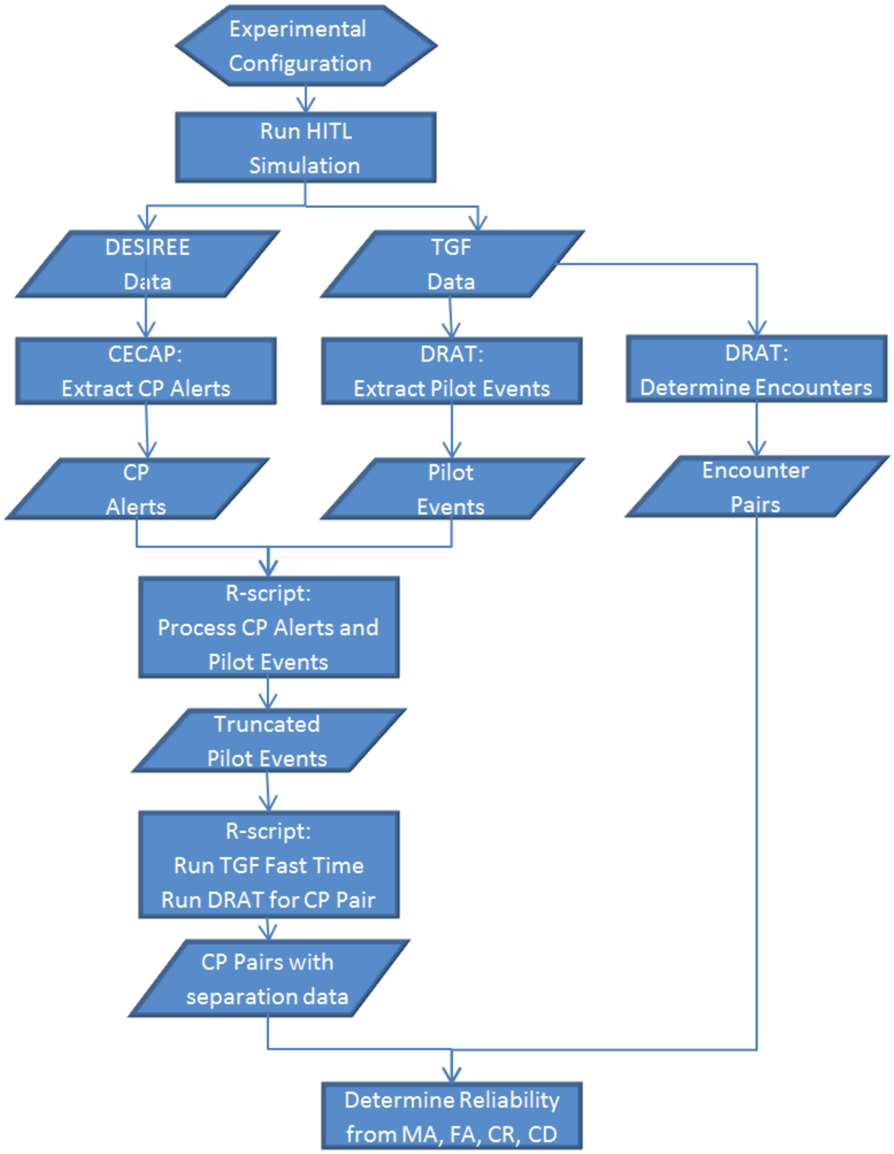

Figure 83. Process and data used to determine CP reliability during the experiment.

4. DISCUSSION

First, it is important to acknowledge that the technical challenges (see Section 2.10) altered the experimental conditions. This requires us to consider each of the primary research questions behind this study and address how the technical challenges impacted our ability to answer the given question. In this section, in light of these issues, we discuss the set of relevant results we were able to obtain, despite the challenges, results that address the questions of interest.

4.1 Minimum CP Alert Reliability

A primary question we intended to address with this experiment was the effect of CP Alert reliability. We had designed the sets of CP adherence bounds to create several levels of CP reliability. Instead, the conformance bounds were always set to 2.5 NM (lateral) by 2.5 NM (longitudinal), with a 10-20 likelihood setting. However, because the Algorithm variation ran mostly as planned, the CP reliability did vary for some conditions, and we were able to assess some effects of the reliability. Most importantly, we could analyze the key question regarding the minimum level of alert reliability that controllers will accept on the R-Side.

Fincannon et al. (2015) showed that the conformance bound and likelihood settings that we ran would produce levels of alert reliability below what controllers currently use in the field. Even at this low level of alert reliability, our observations and data indicated a willingness to use the R-Side CP and indicated evidence of its advantages for subjective preference and controller performance. We cover these effects throughout this section.

The algorithmic enhancements changed how the CP presented false alerts to controllers. The CP presents both red and yellow alerts, and changing the conformance bound settings primarily affected the number of false alerts, which mainly appear as yellow alerts. Given that the conformance bound settings had little to no effect on the number of red alerts, the fact that we showed only red alerts in the algorithmic enhancements condition changed how controllers experienced our manipulations to CP settings. Specifically, when the R-Side only displays red alerts, differences in false alerts should be minimal as conformance bounds change. Because controllers were too busy with traffic to use the RA-Side, a manipulation to CP settings should not have produced a strong effect.

One impact of false alerts in our PSQ data was the effect of Algorithm and CP Location on nuisance alerts. We found a significant increase in reported nuisance alerts, and higher ratings of distraction, when the CP presented alerts on the R-Side with legacy settings. In fact, when the CP was on the R-Side with algorithmic enhancements or was only on the RA-Side, controllers did not report any nuisance alerts. This finding indicates that showing only red CP Alerts on the R-Side—at least if the algorithmic enhancements we studied were in place—would not present a nuisance or distraction to controllers, even at the low reliability levels our participants experienced.

Some findings from our analysis of workload also illustrated the benefit of algorithmic enhancements. The fNIRS analysis indicated that CP algorithmic enhancements were associated with reduced oxygenated hemoglobin levels, indicating a lower level of workload. However, we do not conclude strongly that the algorithmic improvements reduce workload, because of the incomplete nature of our fNIRS data. Furthermore, subjective assessments with the PSQ and WAK did not yield a significant effect of this manipulation.

Regarding CP Alert reliability, some evidence indicates that controllers might have preferred to see more false alerts of situations that were "near-losses" of separation. For example, one controller in the algorithmic enhancements condition ran a vector line to 8 minutes for an aircraft pair and said, "How are you two not red? I'll dump the Delta as soon as I get him." This statement referred to a near-loss of separation at approximately 6 NM, for which the controller expected the CP to provide an alert. Although the CP did report the alert as yellow on the RA-Side, the controller only used or looked at the R-Side and interpreted the situation as a missed alert. This example illustrates an automation trust problem, where the CP functioned properly, but the location of its alerts—where the controller could not readily view them—potentially caused the controller to question its alert reliability.

The discussion of yellow alerts during the post-experiment debriefing was also related to this topic of near-loss of separation. Several controllers stated a preference to see alerts for aircraft that were close but just outside of the 5 NM threshold for loss of separation, which the CP Algorithm used as a cutoff for showing red alerts. In this discussion, the controllers expressed a desire for yellow alerts for near-loss of separation beyond the 5 NM threshold, but there was no consensus on the exact separation distance for which yellow alerts were wanted; controllers' preferences ranged from 5.5 NM to 8 NM. Future research could examine whether the addition of a CP buffer zone permitting the display of near-loss of separation, yellow alerts, improves controller attitudes, workload, and performance.

4.2 CP Location

The Location manipulation, where CP Alerts were either on the RA-Side Only or the R- and RA-Side, was unaffected by the problems with the simulation environment. Therefore, we were able to study and interpret its main effect as planned.

While adding the CP to the R-Side did not reduce workload (see results for WAK and fNIRS), this manipulation appeared to improve ratings on a variety of subjective measures. When the CP was on the R-Side, controllers said that the CP was more useful. The over-the-shoulder observers provided a similar assessment, indicating by their ratings that adding the CP to the R-Side improved the controllers' ability to (a) use the CP in appropriate situations, (b) use the CP in a timely and effective manner, and (c) use the CP overall.

CP availability on the R-Side also appeared to improve controller performance. We examined how controllers corrected suboptimal altitude clearances and found that presenting a CP Alert on the R-Side more than halved the average time to address this problem. We provide more anecdotal discussion of findings in the following section.

Observers provided lower ratings on several dimensions during the condition with CP on the R-Side with algorithmic enhancements. Ratings that decreased in this combination of conditions include (a) preplanning actions, (b) maintaining overall attention and situation awareness, and (c) maintaining attention and situation awareness of aircraft positions. Though controllers appeared to perform better with the CP on the R-Side, these ratings indicated a slight tendency to be reactive to the CP rather than proactively search for conflicts. This finding may seem counterintuitive, but it is consistent with previous research. Endsley and Kaber (1999) noted that higher levels of automation improve performance on control tasks, but reliance on automation also reduces information processing and situation awareness. While the effects in our study were small, they are consistent with this trend. Future research could address this question with regard to the R-Side CP by collecting data that focuses more directly on its impact on attention and situation awareness measures.

4.3 Anecdotes for Conflict Probe with Algorithmic Enhancements

Although many of the objective and subjective findings presented in this report illustrate benefits of implementing the CP on the R-side, there was also qualitative, anecdotal data that could suggest otherwise. This section only presents observations from conditions when the CP was on the R-Side.

In one of the runs with legacy algorithms, a participant expressed frustration about receiving too many CP Alerts, while 11 alerts were in the CPAV. The controller said, "Who are you coming together with? Oh, that guy," and then told one of the observers, "It's almost giving me too many alerts now." In another run, a controller asked the observer about the muted yellow and red alerts and told the observer, "No, I don't like this," followed by, "It makes it totally useless." Examples such as these illustrate that we will need to find a way to provide controllers with necessary information without overloading them with alerts.

Although most controllers thought that muted and yellow alerts created too many alerts, they were also not comfortable when aircraft were passing in close proximity without any alerts. A controller may be comfortable with an aircraft following or passing another aircraft at a distance of 5.5 miles on a clear day with smooth rides, but if turbulence or cloud layers are present, they would not be comfortable without taking action to ensure the aircraft maintain expected trajectories. We observed many examples in the experiment of controllers amending altitudes for an aircraft, issuing vectors to an aircraft, or expediting an aircraft which had a yellow Probe alert in the ACL. Although the aircraft maintained separation, the controller was not confident they would remain separated. A buffer zone may be necessary that alerts controllers and helps to ensure that they maintain positive separation.

In a run with R-Side CP and algorithmic enhancements, a controller created an aircraft conflict while resolving an airspace conflict. A pilot checked in, and the controller gave a 010 heading to go around the SAA, which redirected the aircraft into the path of another aircraft. After the CA activated, the controller then said, "No alert on those two [that is, there had been no CP Alert]; I completely forgot about it." If the CP had presented a buffer zone of alerts, the aircraft pair put into CA status by this action might have already had a (perhaps yellow) CP Alert, the controller would be aware of potential conflicts, and they would have taken an alternate course of action. Further research into an appropriate buffer zone may help controllers avoid this type of error.

In the current design of CP, when multiple alerts are present, there is no indication of which aircraft pairs are associated with each other without additional controller interaction. In its current state in the ACL, the CP Alert only presents number of alerts for each aircraft, but it does not indicate the other aircraft in the conflict. Unless there is only one pair of aircraft alerting with each other, which is rarely the case, the controller must interact with the ACL or GPD and display the trajectories to discover which aircraft are in conflict with each other. All of this is also true for the R-Side CP functionality used in this experiment. In one run with R-Side CP, the controller said aloud, "Who are you coming together with?" The controller could only find this information by interacting with the portal on the FDB. A simpler method to answer the controller's question should be developed. One solution may be to design the R-Side CPAV so that the two aircraft alerted with each other would be side by side—similar to the display of the CA view—and sorted with the most immediate situation at the top. This would eliminate the need for the controller to interact with the display to be aware of which aircraft were in conflict with each other.

Some controllers do not immediately see a benefit to using the CP, and it may take time and experience with the CP to change this position. One controller consistently provided low ratings of CP utility and also stated this opinion during the post-experiment debriefing. Our qualitative review identified several instances where this participant had a red alert on the RA-Side, and displaying these alerts on the R-Side would have improved performance. In one case, the controller said, "Hell, he's clean," and descended an aircraft into another aircraft. This action produced a red alert on the RA-Side that went unnoticed, and several minutes later a CA activated on the R-Side. Though this participant did not see a benefit to the R-Side CP, the function would have avoided the CA.

Other concerns with implementing CP on the R-Side must also be addressed, such as lack of proficiency due to overreliance on automation and how to determine a trainee's ability to detect conflicts. The latter concern was raised by many on-the-job instructors when URET was first introduced. It is very difficult now for RA-position trainees to demonstrate their ability to detect conflicts and for a trainer to evaluate the trainee's ability because conflicts are shown to them well in advance. These trainee weaknesses are not normally discovered now until much later when R-position training starts. CP on the R-position would delay discovery of these proficiency issues even more.

4.4 Biases in Experiment

The participants in this experiment were not the usual population to participate in an ATC HITL study conducted by the Human Factors Branch. Typically, we use controllers that are currently working traffic full-time. Though our participants were all currently certified to work traffic, as mentioned in Section 2.1, they were also members of the ERAM National User Team who spend a good amount of time off the boards. The nature of our sample has both pros and cons. On the one hand, their data might be less generalizable to the controller population. Because the National User Team is charged with very close scrutiny of ATC automation's user interfaces and functions, this population may have been biased toward finding flaws with the system. On the other hand, their familiarity with CP concepts and with ERAM functionality—as compared, for example, to a controller who works traffic full-time at a facility that has only recently begun Full Operational Capability use of ERAM—meant that they required less training and might have been able to provide more insightful opinions regarding the concepts we were studying. We were able to collect a rich set of data from these groups, and their comments during the simulations and debriefing sessions were quite useful. We recommend, however, if logistical considerations permit, that future CP research include, as part of the sample, controllers more similar to the general population and less likely to bring preconceived notions to their participation.

Compared with many HITL experiments that we conduct at the RDHFL, the over-the-shoulder observers had more opportunities to interact with the participants; they were seated at the RA-Side position and were permitted to converse with the participants at all times. Although this factor could result in more accurate assessments of controller performance, it also raises potential issues. Observers were not seen creating inappropriate distractions to the participants and generally focused on observing and rating the participants' performance. However, in a small number of cases, they provided assistance to the participants, despite the point made during the training that the observer was not to act as the participants' RA-Side controllers. As noted in Section 3.5.2, of the commands entered during the analyzed runs, 2.7% were entered from the RA-Side. Though we cannot quantify how many of these the observers initiated, we seldom witnessed participants interacting with the RA-Side console and quite often saw the observers doing so. In addition to operating the RA-Side position as an RA-Side controller would do in a two-person sector team, the observers sometimes pointed out situations for the participant to be aware of and act on, such as descending a flight to meet an SOP. When we were aware of this behavior for a particular event, we either excluded that event from analysis or did not give the controller credit for correct behavior—but without extensive review of video footage, we do not have a complete list of simulation events that might have been affected by observer behavior. However, our conclusion regarding assistive behavior by observers is that it was not excessive and did not appear to be occurring more for particular participants or in particular conditions. We do not believe it compromises our comparisons between conditions or our conclusions. However, in future research, we may return to a more standard RDHFL practice in which observers are slightly physically removed from participants and are more explicitly instructed to refrain from assisting in the ATC tasks.

4.5 Implications of Experiment Complications

The simulation issues that prevented the conformance bounds and likelihood settings from varying between conditions did not, as discussed in Section 4.1, make it impossible to study the main question of required CP Alert reliability for R-Side implementation. However, the situation did compromise our ability to assess more complex relationships between the exact degree of CP reliability and controller behavior, workload, and performance. We used these errors as motivation to revisit our formalized experiment preparation and shakedown process. We faced several issues with developing the simulation environment for this experiment, including the unique challenges of developing the EES-DESIREE interfaces and communications needed to support the first HITL experiment involving this level of integration between the systems.

These challenges cut into the time available to properly dry-run and shake down the scenarios and the environment. They also limited our ability to ensure that the parameter settings we had developed to affect the CP reliability variations were working as desired. Earlier work in the design helped us verify conclusively that the parameter settings were working, but it was only after this verification that a setting change disabled the ability to read in the specified parameters and override the default EES CP settings.

Regardless of these factors, the lesson is that scenario and environment testing, and full shakedown prior to conducting an experiment of this scale, cannot be compromised. If we conduct another experiment involving CP reliability, we have identified steps unique to this type of experiment that we will execute in addition to standard dry runs and shakedowns. These steps include dry runs using test scenarios that we have developed specifically to test certain aspects of the CP algorithms.

Although for many of our analyses we used only 50% of the collected experimental data—the data from the runs that ran mostly as planned—the remainder of the data were not wasted. For example, we used the data from all runs for the altitude clearance analyses, which were primarily concerned with CP Location—a variable that as mentioned above was implemented correctly in all runs. This data provided some of our richest results with regard to potential operational benefits of R-Side CP. Furthermore, we were able to use the fNIRS data from the runs where reliability did not vary as planned. The conditions in which we had chosen to collect fNIRS data, and the factors on which these conditions varied even with the simulation issues, enabled us to test the main effects of our three primary IVs—CP Location, Algorithm, and Noise—on physiologically assessed mental workload. Furthermore, the incorrect implementation of the likelihood setting overrides enabled a more pure test of algorithmic improvements than initially planned. To conserve the number of runs needed, the original experiment plan called for varying both the likelihood function and the presence of modeling enhancements as a manipulation of the Algorithm variable. If the likelihood had been allowed to vary between the two levels of this variable, it would be impossible to conclude whether any observed effects of the algorithmic enhancements were brought about by the modeling enhancements or by the likelihood settings. However, as it turned out, we could investigate the pure effects of the modeling enhancements; in fact, several of our measures did provide evidence of a benefit from these enhancements.

5. RECOMMENDATIONS AND CONCLUSIONS

5.1 Future Research

We can use findings from this HITL experiment to provide recommendations for future research. The following subsections detail specific topics of interest.

5.1.1 Loss of Separation Buffer

Throughout the results and discussion, we noted how controllers benefited from only displaying red alerts on the R-Side, but some controllers stated that there might be utility to displaying a limited number of yellow alerts. Alerts that indicate a near-loss of separation could have helped controllers by reducing the time spent wondering if the CP missed a conflict or by redirecting aircraft that did not require redirection. Controllers occasionally created conflicts by redirecting one aircraft into the flight path of a second aircraft with which it was already nearly in conflict, and presenting alerts for near-loss of separation events might have helped controllers avoid these circumstances. In spite of these potential benefits, controllers did not have any consensus on their recommendations for an appropriate zone; exact distances ranged from 5.5 NM to 8.0 NM. The results of further research into a buffer zone for near-loss of separation alerts could benefit controllers.

5.1.2 Trial Planning

As mentioned above, results from this experiment indicated that controllers occasionally responded to conflicts or potential conflicts by redirecting aircraft into each other's flight path and creating new conflicts. To resolve this issue and minimize "give back" situations, in which a clearance causes a conflict and the controller needs to amend it after issuing it, future research should consider methods of integrating trial planning on the R-Side display.

Based on the experimental conditions in this study, there are several methods of trial planning that future studies could compare and contrast. For example, a future experiment could include a baseline condition using current methods in which controllers would simply use the information that they have in the NAS routes to try and avoid future conflicts. Using CP Alert settings from this HITL experiment, another level of trial planning would involve avoiding conflicts by relying on CP Alerts from (a) a near-loss of separation buffer zone or (b) a post decision period to redirect aircraft. A third level of trial planning could build off of current CP settings by allowing controllers to actively probe a new route for an aircraft before redirecting the aircraft. A fourth option for trial planning could incorporate more automation into the process by examining potential modifications to a flight plan in response to a CP Alert and providing a simplified set of recommendations in probed menus. There are many options for trial planning that build off of the CP for future research.

5.1.3 Electronic Controller to Controller Coordination

The increase in more proactive conflict resolution or airspace avoidance that greater CP use could bring may result in a greater need for controllers to coordinate with each other to maintain positive separation between aircraft. For example, there may be a dynamic SAA or weather event that pilots need to avoid, and the airspace to avoid straddles two or more sectors. Controllers would need to redirect aircraft, but do so in a way that allows a controller in the next sector to expect the redirection. One controller could use point-outs to direct the aircraft around the SAA or weather and use integrated trial planning and coordination functionality to maintain mutual awareness between all controllers that would handle the aircraft. Future research needs to consider methods to support coordination between controllers in these situations.

5.2 Concluding Remarks

In spite of the technical complications, we presented controllers with a level of CP Alert reliability on the R-Side that is below the current standard in the field, so if we display CP Alerts on the R-Side with algorithmic enhancements and current parameter settings, controllers are likely to accept the CP. Results from this experiment illustrated benefits to adding CP Alerts to the R-Side display with algorithmic enhancements, which include improved (a) controller performance, (b) controller perceptions of the CP, and (c) observer ratings of how controllers used the CP. We also identified several areas for future research, including (a) a buffer zone for near-loss of separation alerts, (b) trial planning, and (c) controller to controller coordination.

References

Ayaz, H. (2010). *Functional near-infrared spectroscopy based brain computer interface* (Doctoral dissertation). Philadelphia, PA: Drexel University.

Ayaz, H., Willems, B., Bunce, S., Shewokis, P. A., Izzetoglu, K., Hah, S., . . . Onaral, B. (2010). Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. In T. Marek, W. Karwowski, & V. Rice (Eds.), *Advances in understanding human performance: Neuroergonomics, human factors design, and special populations* (pp. 21–32). CRC Press, Taylor & Francis Group.

Crowell, A., Fabian, A., Young, C. Musialek, B., & Paglione, M. (2011). *Evaluation of prototype enhancements to the En Route Automation Modernization's conflict probe* (DOT/FAA/TC-TN12/3). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Crowell, A., Fabian, A., Young, C., Musialek, B., & Paglione, M. (2012). *Evaluation of prototype enhancements to the En Route Automation Modernization's conflict probe with an updated traffic sample from Chicago* (DOT/FAA/TC-TN12/56). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Crowell, A., & Schnitzer, B. (2013). *Analysis of likelihood parameter changes in the En Route Automation Modernization* (DOT/FAA/TC-TN13/13). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Endsley, M. R., & Kaber, D. (1999). Level of automation effects on performance, situation awareness, and workload in a dynamic control task. *Ergonomics, 42*(3), 462–492. doi:10.1080/001401399185595

Federal Aviation Administration. (2014). *Air Traffic Control* (JO 7110.65V). Retrieved from http://www.faa.gov/regulations_policies/orders_notices/index.cfm/go/document.information/documentID/1023549

Federal Aviation Administration. (2015a). *En Route Automation Modernization (ERAM). Air Traffic Manual (ATM): R-position user manual* (TI 6110.100 EAD300). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Federal Aviation Administration. (2015b). *En Route Automation Modernization (ERAM). Air Traffic Manual (ATM): RA-position user manual* (TI 6110.101 EAD300). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Fincannon, T., Willems, B., & Masalonis, T. (2015). *Evaluation of using conformance bound settings to manipulate the reliability of a conflict probe.* Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center. Manuscript in preparation.

Higgins, J. S., Willems, B., Johnson, D. R., & Zingale, C. M. (2012). *Human factors evaluation of pointing devices used by Air Traffic Controllers: Changes in physical workload and behavior* (DOT/FAA/TC-12/63). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Kell, S., & Chen, Y. (2004). *Initial design for ground delay monitoring capabilities on CAPER* (MP 04W000070). McLean, VA: MITRE Center for Advanced Aviation System Development.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). Elsevier. ISBN: 978-0-12-405888-0. http://dx.doi.org/10.1016/B978-0-12-405888-0.00001-5

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the Third International Workshop on Distributed Statistical Computing* (DSC 2003). Vienna, Austria. Retrieved from https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Preacher, K. J. (2001). *Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence* [Computer software]. Retrieved from http://quantpsy.org/chisq/chisq.htm

R Core Team. (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. ISBN 3-900051-07-0. Retrieved from https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf

Sollenberger, R. L., Stein, E. S., & Gromelski, S. (1997). *The development and evaluation of a behaviorally based rating form for assessing air traffic controller performance* (DOT/FAA/CT-TN96/16). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Sollenberger, R., Willems, B., DiRico, J., Hale, M., & Deshmukh, A. (2010). *Human-in-the-loop investigation of automation requirements for separation management* (DOT/FAA/TC-10/07). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Stein, E. S. (1985). *Air Traffic Controller workload: An examination of workload probe* (DOT/FAA/CT-TN84/24). Atlantic City International Airport, NJ: Federal Aviation Administration Technical Center.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Pearson. ISBN-10: 0-205-45938-2.

Thompson, S. D., & Flavin, J. M. (2006). Surveillance accuracy requirements in support of separation services. *Lincoln Laboratory Journal, 16*(10), 97–122.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society 1*(2), 217–235.

Young, C., Schnitzer, B., DiDonato, L., & Yao, C. (2014). *Evaluation of track monitoring and conflict probe vertical bounds in the En Route Automation Modernization* (FAA-ERAM-2012-0092). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Zingale, C. M., Willems, B., Schulz, K., & Higgins, J. S. (2012). *Separation management (SepMan 2): Human factors evaluations of conflict probe location and format and workstation display alternatives* (DOT/FAA/TC-12/62). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Acronyms

ACL	Aircraft List
ARTCC	Air Route Traffic Control Center
ATC	Air Traffic Control
CA	Conflict Alert
CAK	Akron-Canton Regional Airport
CAPER	CAASD Analysis Platform for En Route
CHI	Computer-Human Interface
CMH	Port Columbus International Airport
CP	Conflict Probe
CPAV	Conflict Probe Alert View
CPC	Certified Professional Controller
CVG	Cincinnati/Northern Kentucky International Airport
DESIREE	Distributed Environment for Simulation, Rapid Engineering, and Experimentation
DRAT	Data Reduction and Analysis Tool
EES	ERAM Evaluation System
ERAM	En Route Automation Modernization
FAA	Federal Aviation Administration
FDB	Full Data Block
fNIRS	functional Near-Infrared Spectroscopy
GPD	Graphic Plan Display
HDI	High-Density Interval
HITL	Human-In-The-Loop
IVs	Independent Variables
JAGS	Just Another Gibbs Sampler
LOA	Letter of Agreement
MAP	Monitor Alert Parameter
NAS	National Airspace System
NM	Nautical Mile
PSQ	Post-Scenario Questionnaire
PTT	Push-To-Talk
RA-Side	Radar Associate Side
RDHFL	Research Development and Human Factors Laboratory

ROPE	Region of Practical Equivalence
R-Side	Radar Side
SAA	Special Activity Airspace
SDRR	Simulation Driver and Radar Recorder
SME	Subject Matter Expert
SOP	Standard Operating Procedure
TGF	Target Generation Facility
URET	User Request Evaluation Tool
VSCS	Voice Switching and Control System
WAK	Workload Assessment Keypad
WJHTC	William J. Hughes Technical Center
ZAU	Chicago ARTCC
ZDC	Washington ARTCC
ZID	Indianapolis ARTCC

Appendix A: Informed Consent Form

Informed Consent Statement

I, ______________________________, understand that this study, entitled "Separation Management Human-in-the-Loop Simulation # 3: Evaluations of Conflict Probe Accuracy Requirements, Sector Enhancements Displays, Conflict Probe Algorithmic Improvements, and Alerting Parameters", is sponsored by the Federal Aviation Administration (FAA).

Nature and Purpose:

I have been recruited to volunteer as a participant in this project. The simulation will investigate: (a) required accuracy levels for Conflict Probe implementation on the R-Side, (b) proposed R-Side and RA-Side display enhancements, (c) Conflict Probe algorithmic improvements, and (d) alerting parameter schemes for potential conflicts. I understand that I will be assigned to work the R-Side and/or RA-Side during the simulation. I will be asked to wear functional near-infrared spectroscopy (fNIRS) equipment to record brain oxygenation and/or eye tracking equipment to record what I looking at on the interface and assess my workload. The results of this study will be used to evaluate the benefits and feasibility of integrating the aforementioned Conflict Probe components into the future en route environment.

Study Procedures:

Eight (8) certified professional controllers (CPCs) from the ERAM Subject Matter Expert (SME) group will participate in the simulation. Four participants will participate at one time. They will spend four days at the lab over a one-week period, spanning Saturday and Sunday. At the start of the simulation, the participants will be assigned to work as either an R-Side or RA-Side controller for conditions that require R-Side/RA-Side teams. Other conditions will require that each participant work as an R-Side alone. Participants will work from about 8:00 AM to about 4:30 PM every day, with a lunch break and at least two rest breaks. At the end of each day, there will be a group meeting during which participants can ask questions and discuss the day's simulations. The first morning, Day 1, will consist of an initial briefing outlining the nature of the study and what participants can expect. Participants will be introduced to the simulator equipment. For the remainder of that day, the participants will be trained on the novel tools and procedures being evaluated and will complete 45 to 60 minute training simulation runs. On Day 2, the participants will complete approximately three hours of additional training and practice, which will be followed by four test runs. The remaining two days will have up to six test runs.

Participants will provide real-time workload ratings at two-minute intervals during all runs. During practice and test runs, participants will wear functional near-infrared spectroscopy (fNIRS) or eye tracking equipment. The simulator will automatically record system operations. All simulation runs will be audio and video recorded. After each test run, participants will complete questionnaires related to their workload, situation awareness, and other aspects of the simulation run.

After the final test run, participants and researchers will conduct a final debriefing session to share questions, comments, and feedback.

Anonymity and Confidentiality:

My participation in this simulation is strictly confidential. Any information I provide will remain anonymous; no individual names or identities will be associated with the data or released in any reports.

Benefits:

I understand that the only benefit to me is that I will be able to provide valuable feedback and insight into the effectiveness of potential ATC tools and procedures. My contribution will help the FAA to determine the benefits and feasibility of these modifications.

Participant Responsibilities:

I am aware that to participate in this study I must be a certified professional controller who is qualified at my facility and holds a current medical certificate. I must have normal or corrected-to-normal (20/20) vision. I will control simulated traffic and answer the questions asked during the study to the best of my ability. I will not discuss the content of the study with anyone until the study is completed.

Participant Assurances:

I understand that my participation in this study is completely voluntary and I can withdraw at any time without penalty. I also understand that the researchers in this study may terminate my participation if they feel this to be in my best interest. I understand that if new findings develop during the course of this study that may relate to my decision to continue participation, I will be informed. I have not given up any of my legal rights or released any individual or institution from liability for negligence.

The research team has adequately answered all the questions I have asked about this study, my participation, and the procedures involved. I understand that Ben Willems or another member of the research team will be available to answer any questions concerning procedures throughout this study. If I have questions about this study or need to report any adverse effects from research procedures, I will contact Ben Willems at (609) 485-4191.

Discomfort and Risks:

I understand that I will not be exposed to any known risks or intrusive measurement techniques. Some participants may experience discomfort from the head-mounted fNIRS and eye tracking equipment. I agree to immediately report any injury or suspected adverse effect to Ben Willems at (609) 485-4191.

Signature Lines:

I have read this informed consent form. I understand its contents, and I freely consent to participate in this study under the conditions described. I understand that, if I want to, I may have a copy of this form.

Participant:____ Date:_____________

Investigator:____ Date:_____________

Witness:____ Date:_____________

Appendix B: Biographical Questionnaire

Biographical Questionnaire

This questionnaire is designed to obtain information about your background and experience as a certified professional controller (CPC). Researchers will only use this information to describe the participants in this study as a group. Your identity will remain anonymous.

1. What is your **gender**?	○ Male	○ Female

2. What is your **age**?	_____ **years**
3. How long have you worked as **an Air Traffic Controller (include both FAA and military experience)**?	_____ **years** _____ **months**
4. How long have you worked as a **CPC for the FAA**?	_____ **years** _____ **months**
5. How long have you **actively controlled traffic** in the **en route** environment?	_____ **years** _____ **months**
6. How long have you **actively controlled traffic** in the **terminal** environment?	_____ **years** _____ **months**
7. How many of the **past 12 months** have you actively controlled traffic?	_____ **months**
8. How long have you been **using ERAM operationally**?	_____ **years** _____ **months**
9. When did you last receive **ERAM training**?	_____ **/** ________ **(month/year)**

10. Rate your current **skill as a CPC**.	Not Skilled	①②③④⑤⑥⑦⑧⑨⑩	Extremely Skilled
11. Rate your **level of motivation** to participate in this study.	Not Motivated	①②③④⑤⑥⑦⑧⑨⑩	Extremely Motivated

Appendix C: Post-Scenario Questionnaire

Participant # ___________ **Date ___________**

Post-Scenario Questionnaire

Part 1 – Overall Performance, Workload, Situation Awareness, and Simulation Ratings

Answer the following questions based upon your experience in the scenario just completed. Fill in one circle to indicate your response to each item.

1. Rate the **overall difficulty** of this scenario.	Extremely Difficult	①②③④⑤⑥⑦⑧⑨⑩	Extremely Easy
2. Rate your **overall level of ATC performance**.	Poor	①②③④⑤⑥⑦⑧⑨⑩	Excellent
3. Rate your level of performance in **resolving conflicts**.	Poor	①②③④⑤⑥⑦⑧⑨⑩	Excellent
4. Rate your level of performance in **all other tasks**.	Poor	①②③④⑤⑥⑦⑧⑨⑩	Excellent
5. Rate your **overall workload**.	Poor	①②③④⑤⑥⑦⑧⑨⑩	Excellent
6. Rate your **overall level of situation awareness**.	Poor	①②③④⑤⑥⑦⑧⑨⑩	Excellent
7. Rate the **performance of the simulation pilots** in terms of their responses to control instructions and providing callbacks.	Poor	①②③④⑤⑥⑦⑧⑨⑩	Excellent

Write number responses on the lines provided.

8. What was the **average number of aircraft** in the sector?	_________ aircraft
9. What was the **largest number of aircraft** in the sector at one time?	_________ aircraft

10. What aspects of this scenario were easiest to work with? Why?

11. What aspects of this scenario were hardest to work with? Why?

12. Please write any additional comments you have about your experience in this scenario.

Participant # ___________ **Date** ___________

Part 2: Conflict Probe

The following questions ask you to estimate the performance of Conflict Probe in the scenario just completed. Write number responses on the lines provided.

1. Of **all actual/probable conflicts,** what percentage were **never alerted** by Conflict Probe?	________%
2. Of **all Conflict Probe alerts**, what percentage were false alerts?	________%
3. What was the **total** number of **alerts**?	________ alerts
4. Of **all accurate Conflict Probe alerts**, what percentage would you describe as **nuisance alerts**, that is, accurate but of no use to you and you would prefer not to have them?	________%
5. What caused alerts in this scenario to be nuisance alerts, if there were any? (For example: too early; too late; accurate information but separation distance too high—or too low—for the alert to be useful?)	
6. When Conflict Probe gave **accurate alerts**, what was the average **warning time** before the conflict would have occurred?	______ minutes ______ seconds

Answer the following questions based upon your experience in the scenario just completed. Fill in one circle to indicate your response to each item.

7. How **accurate** was Conflict Probe during the scenario?	Not at all accurate	①②③④⑤⑥⑦⑧⑨⑩	Extremely accurate
8. How **useful** was Conflict Probe during the scenario?	Not at all useful	①②③④⑤⑥⑦⑧⑨⑩	Extremely useful
9. How **distracting** was Conflict Probe during the scenario?	Not at all distracting	①②③④⑤⑥⑦⑧⑨⑩	Extremely distracting
10. Were you prepared to trust the Conflict Probe during the scenario?		Yes / No	
11. How much confidence did you have in the detection accuracy of the Conflict Probe during the scenario?	Not at all confident	①②③④⑤⑥⑦⑧⑨⑩	Extremely confident
12. How much confidence did you have in your ability to detect conflicts in the scenario?	Not at all confident	①②③④⑤⑥⑦⑧⑨⑩	Extremely confident
13. How much confidence did you have that you and the Conflict Probe together could detect conflicts in the scenario?	Not at all confident	①②③④⑤⑥⑦⑧⑨⑩	Extremely confident

Participant # ___________ **Date** ___________

14. For the items below, please indicate how you detected and resolved conflicts during **<u>this scenario</u>**.
____

A. I responded to a conflict without question after I received an alert from the Conflict Probe. ☐ ☐ ☐ ☐ ☐ Always Often Sometimes Rarely Never
B. I noticed alerts from the Conflict Probe and believed that it was correct, but I checked other data before I responded to the conflicts. ☐ ☐ ☐ ☐ ☐ Always Often Sometimes Rarely Never
C. I noticed alerts from the Conflict Probe and did not know if they were correct, so I checked other data before responding to the conflicts. ☐ ☐ ☐ ☐ ☐ Always Often Sometimes Rarely Never
D. I detected conflicts on my own and did not know if I was correct, so I checked the alerts from the Conflict Probe before responding to the conflicts. ☐ ☐ ☐ ☐ ☐ Always Often Sometimes Rarely Never
E. I detected conflicts on my own and believed that I was correct, but I checked the alerts from the Conflict Probe before responding to the conflicts. ☐ ☐ ☐ ☐ ☐ Always Often Sometimes Rarely Never
F. I detected and resolved conflicts without using the Conflict Probe. ☐ ☐ ☐ ☐ ☐ Always Often Sometimes Rarely Never

15. Please write any additional comments you have about your experience working with Conflict Probe in this scenario.

Appendix D: Exit Questionnaire

Participant # ___________ **Date** ___________

Exit Questionnaire

Part 1 – Simulation Realism and Research Apparatus Ratings

Please respond to each of the following items based upon your overall experience in the simulation. Fill in one circle to indicate your response to each item.

1. Rate the **overall** realism of the simulation.	Not at all Realistic	①②③④⑤⑥⑦⑧⑨⑩	Extremely Realistic
2. Rate the **realism of the simulation hardware** compared to actual equipment.	Not at all Realistic	①②③④⑤⑥⑦⑧⑨⑩	Extremely Realistic
3. Rate the **realism of the simulation software** compared to actual equipment.	Not at all Realistic	①②③④⑤⑥⑦⑧⑨⑩	Extremely Realistic
4. Rate the **realism of the simulation traffic scenarios** compared to actual NAS traffic.	Not at all Realistic	①②③④⑤⑥⑦⑧⑨⑩	Extremely Realistic
5. To what extent did the **WAK online workload rating technique interfere** with your ATC performance?	Not At All	①②③④⑤⑥⑦⑧⑨⑩	A Great Deal
6. How effective was the **training** provided?	Not At All Effective	①②③④⑤⑥⑦⑧⑨⑩	Extremely Effective
7. ***Answer only if you wore the fNIRS***: To what extent did the fNIRS interfere with your ATC performance?	Not At All	①②③④⑤⑥⑦⑧⑨⑩	A Great Deal
8. ***Answer only if you used the eye tracking equipment***: To what extent did the eye tracking equipment interfere with your ATC performance?	Not At All	①②③④⑤⑥⑦⑧⑨⑩	A Great Deal

9. Please write any additional comments about the simulation that you would like us to know about.

Participant # ___________ **Date ___________**

Part 2 – Conflict Probe

Please respond to each of the following items based upon your overall experience with the Conflict Probe **<u>in the field with standard, legacy settings</u>**.

Write number responses on the lines provided.

1. Of **all actual/probable conflicts in the field <u>with standard, legacy settings</u>,** what percentage are **never alerted** by Conflict Probe?	________%
2. Of **all Conflict Probe alerts in the field <u>with standard, legacy settings</u>**, what percentage are false alerts?	________%
3. Of **all accurate Conflict Probe alerts in the field <u>with standard, legacy settings</u>**, what percentage would you describe as **nuisance alerts**?	________%
4. When Conflict Probe gives **accurate alerts in the field <u>with standard, legacy settings</u>**, what is the average **warning time** before the conflict would have occurred?	______ minutes ______ seconds

Fill in <u>one</u> circle to indicate your response to each item.

5. How **accurate** is Conflict Probe?	Not at all accurate	①②③④⑤⑥⑦⑧⑨⑩	Extremely accurate
6. How **useful** is Conflict Probe when you are working in a team at the RA-Side?	Not at all useful	①②③④⑤⑥⑦⑧⑨⑩	Extremely useful
7. How **useful** is Conflict Probe when you are working in a team at the R-Side?	Not at all useful	①②③④⑤⑥⑦⑧⑨⑩	Extremely useful
8. How **useful** is Conflict Probe when you are working alone at the R-Side?	Not at all useful	①②③④⑤⑥⑦⑧⑨⑩	Extremely useful
9. If you were working in a team at the R-Side, how **distracting** would it be to have Conflict Probe displayed on the R-Side?	Not at all distracting	①②③④⑤⑥⑦⑧⑨⑩	Extremely distracting
10. If you were working alone at the R-Side, how **distracting** would it be to have Conflict Probe displayed on the R-Side?	Not at all distracting	①②③④⑤⑥⑦⑧⑨⑩	Extremely distracting

11. Please write any additional comments you have about your experience working with Conflict Probe in the field.

Participant # ___________ **Date** ___________

How did the following Conflict Probe features affect your performance in this study? Fill in <u>one</u> circle to indicate your response to each item.

12. Conflict Probe alerts in the Conflict Alert List when they were correct	Hurt greatly	①②③④⑤⑥⑦⑧⑨⑩	Help greatly
13. Conflict Probe alerts in aircraft data blocks when they were correct	Hurt greatly	①②③④⑤⑥⑦⑧⑨⑩	Help greatly
14. Increment/decrement function to filter alerts based on time	Hurt greatly	①②③④⑤⑥⑦⑧⑨⑩	Help greatly
15. The specific values of 3, 5, 10, 15, & +20 minutes used in the filtering capability	Hurt greatly	①②③④⑤⑥⑦⑧⑨⑩	Help greatly
16. The fact that an alert portal did not appear until there was a red alert for the R-Side	Hurt greatly	①②③④⑤⑥⑦⑧⑨⑩	Help greatly
17. The fact that yellow alerts were only presented on the RA-Side, not on the R-Side	Hurt greatly	①②③④⑤⑥⑦⑧⑨⑩	Help greatly

18. Please write any additional comments you have about the Conflict Probe functionality and information tested in the simulation.

Appendix E: Over-the-Shoulder Rating Form

Participant # ___________ **Date ___________**

Over-the-Shoulder Rating Form

Please evaluate the effectiveness of the controllers. Please write down observations and make preliminary ratings during the course of the scenario. However, please wait until the scenario is finished before making your final ratings. The observations you make do not need to be restricted to the performance areas covered in this form and may include other areas that you think are important. Also, please write down any comments that may improve this evaluation form. Your identity will remain anonymous, so do not write your name on the form.

Rating	Label Description
1	Minimally Effective
2	
3	
4	Moderately Effective
5	
6	
7	
8	Extremely Effective

Maintaining Safe and Efficient Traffic Flow

1. Maintaining Separation and Resolving Potential Conflicts - using control instructions that maintain safe aircraft separation - detecting and resolving impending conflicts early	1	2	3	4	5	6	7	8
2. Sequencing Arrival and Departure Aircraft Efficiently - using efficient and orderly spacing techniques for arrival and departure aircraft - maintaining safe arrival and departure intervals that minimize delays	1	2	3	4	5	6	7	8
3. Using Control Instructions Effectively - providing accurate navigational assistance to pilots - avoiding clearances that result in the need for additional instructions to handle aircraft completely - avoiding excessive vectoring or over-controlling	1	2	3	4	5	6	7	8
4. Overall Safe and Efficient Traffic Flow	1	2	3	4	5	6	7	8

Maintaining Attention and Situation Awareness

5. Maintaining Awareness of Aircraft Positions - avoiding fixation on one area of the radar scope when other areas need attention - using scanning patterns that monitor all aircraft on the radar scope	1	2	3	4	5	6	7	8
6. Ensuring Positive Control	1	2	3	4	5	6	7	8
7. Detecting Pilot Deviations from Control Instructions - ensuring that pilots follow assigned clearances correctly - correcting pilot deviations in a timely manner	1	2	3	4	5	6	7	8
8. Correcting Own Errors in a Timely Manner	1	2	3	4	5	6	7	8
9. Overall Attention and Situation Awareness	1	2	3	4	5	6	7	8

Participant # ___________ **Date** ___________

Prioritizing

10. Taking Actions in an Appropriate Order of Importance - resolving situations that need immediate attention before handling low priority tasks - issuing control instructions in a prioritized, structured, and timely manner	1	2	3	4	5	6	7	8
11. Preplanning Control Actions - scanning adjacent sectors to plan for inbound traffic	1	2	3	4	5	6	7	8
12. Handling Control Tasks for Several Aircraft - shifting control tasks between several aircraft when necessary - avoiding delays in communications while thinking or planning control actions	1	2	3	4	5	6	7	8
13. Overall Prioritizing	1	2	3	4	5	6	7	8

Providing Control Information

14. Providing Essential ATC Information - providing mandatory services and advisories to pilots in a timely manner - exchanging essential information	1	2	3	4	5	6	7	8
15. Providing Additional ATC Information - providing additional services when workload is not a factor - exchanging additional information	1	2	3	4	5	6	7	8
16. Overall Providing Control Information	1	2	3	4	5	6	7	8

Technical Knowledge

17. Showing Knowledge of LOAs and SOPs - controlling traffic as depicted in current LOAs and SOPs - performing handoff procedures correctly	1	2	3	4	5	6	7	8
18. Showing Knowledge of Aircraft Capabilities and Limitations - avoiding clearances that are beyond aircraft performance parameters - recognizing the need for speed restrictions and wake turbulence separation	1	2	3	4	5	6	7	8
19. Overall Technical Knowledge	1	2	3	4	5	6	7	8

Voice Communications

20. Using Proper Phraseology - using words and phrases specified in JO 7110.65S - using proper phraseology that is appropriate for the situation - avoiding the use of excessive verbiage	1	2	3	4	5	6	7	8
21. Communicating Clearly and Efficiently - speaking at the proper volume and rate for pilots to understand - speaking fluently while scanning or performing other tasks - clearance delivery is complete, correct and timely - providing complete information in each clearance	1	2	3	4	5	6	7	8
22. Listening for pilot readbacks and requests - correcting pilot readback errors - acknowledging pilot or other controller requests promptly - processing requests correctly in a timely manner	1	2	3	4	5	6	7	8
23. Overall Voice Communication	1	2	3	4	5	6	7	8

Participant # ___________ **Date** ___________

Conflict Probes

24. Using Conflict Probes in appropriate situations	1	2	3	4	5	6	7	8
25. Taking full advantage of Conflict Probes in a timely and efficient manner	1	2	3	4	5	6	7	8
26. Overall Conflict Probes use	1	2	3	4	5	6	7	8

Participant # __________ **Date __________**

NOTES

Maintaining Safe and Efficient Traffic Flow

1. Maintaining Separation and Resolving Potential Conflicts

2. Sequencing Arrival and Departure Aircraft Efficiently

3. Using Control Instructions Effectively

4. Overall Safe and Efficient Traffic Flow

Maintaining Attention and Situation Awareness

5. Maintaining Awareness of Aircraft Positions

6. Ensuring Positive Control

7. Detecting Pilot Deviations from Control Instructions

8. Correcting Own Errors in a Timely Manner

9. Overall Attention and Situation Awareness

Prioritizing

10. Taking Actions in an Appropriate Order of Importance

11. Preplanning Control Actions

12. Handling Control Tasks for Several Aircraft

13. Overall Prioritizing

Participant # ___________ **Date ___________**

Providing Control Information
14. Providing Essential ATC Information

15. Providing Additional ATC Information

16. Overall Providing Control Information

Technical Knowledge
17. Showing Knowledge of LOAs and SOPs

18. Showing Knowledge of Aircraft Capabilities and Limitations

19. Overall Technical Knowledge

Voice Communications
20. Using Proper Phraseology

21. Communicating Clearly and Efficiently

22. Listening for pilot readbacks and requests

23. Overall Voice Communication

Conflict Probes
24. Using Conflict Probes in appropriate situations

25. Taking full advantage of Conflict Probes in a timely and efficient manner

26. Overall Conflict Probes use

General Notes:

Appendix F: WAK Anchors

WAK Anchors

Tasks Affected		Rat-ing	Description
Non-Critical ATC Tasks	Examples: Separating data blocks, providing advisories, updating fourth line	1	Very easy, plenty of spare time, you have time to chit-chat
		2	Less spare time left
		3	Even less spare time left, workload may start to affect some non-critical tasks
		4	Some non-critical tasks are left unfinished
		5	Many non-critical tasks are left unfinished
Critical ATC Tasks	Examples: Coordination, spacing, making handoffs, maintaining separation	6	Workload may start to affect some critical tasks
		7	More critical tasks may become affected
		8	It is now likely that some critical tasks are left unfinished
		9	No longer manageable, safety is affected
		10	All critical ATC tasks are affected, you are only focused on keeping aircraft separated

Appendix G: Intended and Actual Parameter Settings in Each of the 12 Conditions

Intended and Actual Parameter Settings

Condition	Intended Settings	Actual Settings	Match
A-R1	CB: 2.5 (Lat) / 2.5 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	√
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: R & RA	POS: R & RA	√
	SN: NO	SN: NO	√
A-R2	CB: 2.5 (Lat) / 2.0 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: R & RA	POS: R & RA	√
	SN: NO	SN: NO	√
A-R3	CB: 1.0 (Lat) / 2.5 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: R & RA	POS: R & RA	√
	SN: NO	SN: NO	√
A-R4	CB: 1.0 (Lat) / 1.25 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: R & RA	POS: R & RA	√
	SN: NO	SN: NO	√
A-R5	CB: 0.1 (Lat) / 0.1 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: R & RA	POS: R & RA	√
	SN: NO	SN: NO	√
A-RA2	CB: 2.5 (Lat) / 2.0 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: RA Only	POS: RA Only	√
	SN: NO	SN: NO	√
A-RA3	CB: 1.0 (Lat) / 2.5 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: RA Only	POS: RA Only	√
	SN: NO	SN: NO	√
A-RA5	CB: 0.1 (Lat) / 0.1 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: RA Only	POS: RA Only	√
	SN: NO	SN: NO	√

Condition	Intended Settings	Actual Settings	Match
B-R1	CB: 2.5 (Lat) / 2.0 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 10-20	LK: 10-20	√
	ALG: NO	ALG: NO	√
	POS: R & RA	POS: R & RA	√
	SN: YES	SN: YES	√
B-R2	CB: 2.5 (Lat) / 2.0 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: R & RA	POS: R & RA	√
	SN: YES	SN: YES	√
B-RA1	CB: 2.5 (Lat) / 2.0 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 10-20	LK: 10-20	√
	ALG: NO	ALG: NO	√
	POS: RA Only	POS: RA Only	√
	SN: YES	SN: YES	√
B-RA2	CB: 2.5 (Lat) / 2.0 (Long)	CB: 2.5 (Lat) / 2.5 (Long)	X
	LK: 4-8-20	LK: 10-20	X
	ALG: YES	ALG: YES	√
	POS: RA Only	POS: RA Only	√
	SN: YES	SN: YES	√

Note. CB = Conformance Bounds; LK = Likelihood Function; ALG = Algorithmic Enhancements; POS = Position of the Conflict Probe; SN = Surveillance Noise.

Appendix H: Detailed WAK Results

Table H1. Mean and (*SD*) WAK Ratings, for All Analyzed Conditions

Condition	Mean and (*SD*) of Rating
R & RA/Enhancements/No Noise	4.50 (2.17)
R & RA/Legacy/Noise	4.33 (2.43)
R & RA/Enhancements/Noise	4.28 (2.49)
R & RA Overall	4.37 (2.37)
RA Only/Enhancements/No Noise	4.57 (2.74)
RA Only/Legacy/Noise	4.17 (2.48)
RA Only/Enhancements/Noise	4.15 (2.51)
RA Only Overall	4.30 (2.58)

The overall means for the three R & RA-Side conditions, vs. the three RA Only conditions, and for the subsets of conditions used in each analysis on WAK Ratings, appear in Tables H2 through H4. Each table is followed by the results of the regression analysis for the effects addressed by the given subset of conditions.

Table H2. Mean and (*SD*) WAK Ratings, Overall by Location

Condition	Mean and (*SD*) of Rating
R & RA Overall	4.37 (2.37)
RA Only Overall	4.30 (2.58)

The effect of Location was not significant, $F(1, 46.09) = 0.40, p < .54$.

Table H3. Mean and (*SD*) WAK Ratings, by Location and Algorithm

Condition	Mean and (*SD*) of Rating
R & RA/Legacy/Noise	4.33 (2.43)
R & RA/Enhancements/Noise	4.28 (2.49)
RA Only/Legacy/Noise	4.17 (2.48)
RA Only/Enhancements/Noise	4.15 (2.51)

The main effect of Location was not significant, $F(1, 28.1) = 0.55$, $p < .47$. The main effect of Algorithm was not significant, $F(1, 28.02) = 0.069$, $p < .80$. The interaction of Location and Algorithm was also not significant, $F(1, 28.02) = 0.092$, $p < .77$.

Table H4. Mean and (*SD*) WAK Ratings, by Location and Noise

Condition	Mean and (*SD*) of Rating
R & RA/Enhancements/No Noise	4.50 (2.17)
R & RA/Enhancements/Noise	4.28 (2.49)
RA Only/Enhancements/No Noise	4.57 (2.74)
RA Only/Enhancements/Noise	4.15 (2.51)

The main effect of Location was not significant, $F(1, 28.02) = 0.056$, $p < .82$. The main effect of Noise was not significant, $F(1, 28.02) = 0.36$, $p < .56$. The interaction of Location and Algorithm was also not significant, $F(1, 28.01) = 0.065$, $p < .81$.

We also computed the overall percentage of ratings that were greater than or equal to 6, where a 6 rating indicates a perception that critical tasks are beginning to be affected, and successively higher ratings represent greater effects on critical tasks. See Table H5 for the percentages of these high ratings in individual conditions. We noted no operationally meaningful differences between any conditions of interest.

Table H5. Percent of WAK Ratings (greater than or equal to 6) for All Compared Conditions

Condition	Percent of Ratings ≥ 6
R & RA/Enhancements/No Noise	32.59%
R & RA/Legacy/Noise	31.71%
R & RA/Enhancements/Noise	34.31%
R & RA Overall	32.91%
RA Only/Enhancements/No Noise	32.37%
RA Only/Legacy/Noise	31.62%
RA Only/Enhancements/Noise	30.88%
RA Only Overall	31.63%

Appendix I: Detailed Results on fNIRS

Table I1 shows the mean and standard deviation of the oxygenated hemoglobin levels for the set of conditions for which we collected fNIRS data

Table I1. Mean and (*SD*) Oxygenated Hemoglobin Levels, Relative to Baseline, for All Analyzed Conditions

Condition	Mean and (*SD*) Oxygenated Hemoglobin Levels (µmol/L)
R & RA/Enhancements/No Noise (A-R1)	1.88 (1.21)
R & RA/Enhancements/No Noise (A-R3)	1.54 (1.84)
R & RA/Enhancement/No Noise (A-R5)	0.96 (2.04)
RA Only/Enhancements/No Noise (A-RA5)	1.61 (1.22)
R & RA/Legacy/Noise (B-R1)	1.94 (1.82)
R & RA/Enhancements/Noise (B-R2)	1.50 (1.26)

We include the above table for a general reference. The three tables that follow present the means for the subsets of the data we analyzed to test each of the three main effects. Each table is followed by the results of the regression analysis for the effects addressed by the given subset of conditions.

Table I2. Mean and (*SD*) Oxygenated Hemoglobin Levels, Relative to Baseline, by Location

Condition	Mean and (*SD*) Oxygenated Hemoglobin Levels (µmol/L)
R & RA Overall (A-R1, AR-3, AR-5)	1.44 (1.79)
RA Only Overall (A-RA5)	1.61 (1.22)

Note. All conditions analyzed here had Algorithmic Improvements and No Noise.

The effect of Location was not significant, $F(1, 515.4) = 1.26, p < .27$.

Table I3. Mean and (*SD*) Oxygenated Hemoglobin Levels, Relative to Baseline, by Algorithm

Condition	Mean and (*SD*) Oxygenated Hemoglobin Levels (μmol/L)
R & RA/Legacy/Noise	1.94 (1.82)
R & RA/Enhancements/Noise	1.50 (1.26)

As reported in Section 3.2, the main effect of Algorithm was significant, $F(1, 229.6) = 4.07$, $p < .045$. This finding indicates higher mental workload with the Legacy algorithms.

Table I4. Mean and (*SD*) Oxygenated Hemoglobin Levels, Relative to Baseline, by Noise

Condition	Mean and (*SD*) Oxygenated Hemoglobin Levels (μmol/L)
No Noise Overall (A-R1, AR-3, AR-5)	1.44 (1.79)
Noise Overall (BR-2)	1.50 (1.26)

Note. All conditions analyzed here had Algorithmic Improvements and R-Side CP Location.

As reported in Section 3.2, the main effect of Noise exhibited a trend toward significance, $F(1, 502) = 2.85$, $p < .10$. The analysis suggests higher mental workload with Surveillance Noise.

Appendix J: Tables of Bayesian Analyses

Table J1. Summary of Bayesian Output for Algorithm by Location Analysis of the PSQ

Question	Mean (SD)				% within the ROPE		
	Legacy/ R & RA	Legacy/ RA	Enhancements /R & RA	Enhancements /RA	Algorithm	Location	Interaction
Rate the overall difficulty of this scenario.	6.22 (2.40)	6.05 (2.07)	5.43 (2.27)	5.81 (2.16)	42.3%	51.1%	38.8%
Rate your overall level of ATC performance.	7.88 (0.77)	8.02 (1.11)	7.63 (1.74)	7.65 (1.13)	64.9%	75.9%	57.2%
Rate your level of performance in resolving conflicts.	8.54 (1.25)	8.14 (1.75)	8.12 (2.01)	8.30 (1.11)	63.4%	62.7%	43.2%
Rate your level of performance in all other tasks.	8.00 (0.85)	8.09 (1.26)	7.89 (0.75)	7.99 (0.85)	86.9%	86.9%	73.8%
Rate your overall workload.	7.93 (1.19)	7.65 (1.05)	7.33 (1.25)	7.20 (1.08)	48.1%	74.5%	63.0%
Rate your overall level of situation awareness.	7.96 (0.96)	7.94 (1.61)	7.59 (2.10)	8.15 (1.20)	67.4%	61.8%	44.4%
Rate the performance of the simulation pilots in terms of their responses to control instructions and providing callbacks.	8.45 (2.57)	8.79 (1.79)	8.54 (1.77)	7.33 (1.98)	34.5%	43.9%	21.4%
What was the average number of aircraft in the sector?	16.02 (20.17)	16.21 (20.23)	15.92 (20.31)	16.49 (19.42)	31.6%	31.5%	28.2%
What was the largest number of aircraft in the sector at one time?	20.59 (24.98)	22.02 (25.95)	20.66 (25.29)	22.00 (25.25)	29.9%	17.7%	35.2%
Of all actual/probable conflicts, what percentage were never alerted by Conflict Probe?	4.96 (9.38)	4.25 (6.61)	5.60 (7.77)	5.12 (9.07)	13.7%	13.8%	13.7%
Of all Conflict Probe alerts, what percentage were false alerts?	0.00 (3.21)	0.00 (3.20)	0.00 (3.20)	0.00 (3.20)	100.0%	100.0%	100.0%
What was the total number of alerts?	14.94 (8.39)	16.59 (11.43)	15.02 (8.40)	15.50 (13.01)	0.3%	0.3%	0.3%
Of all accurate Conflict Probe alerts, what percentage would you describe as nuisance alerts, that is, accurate but of no use to you and you would prefer not to have them?	13.64 (8.22)	0.18 (0.72)	0.00 (0.00)	0.00 (0.01)	0.1%	0.2%	0.1%
When Conflict Probe gave accurate alerts, what was the average warning time before the conflict would have occurred?	451.56 (162.45)	515.55 (249.09)	450.09 (209.49)	485.14 (220.22)	0.6%	0.5%	0.6%
How accurate was Conflict Probe during the scenario?	6.73 (2.65)	6.50 (2.45)	7.34 (2.08)	6.17 (3.09)	40.7%	31.8%	29.3%
How useful was Conflict Probe during the scenario?	6.30 (2.68)	4.51 (3.84)	7.69 (2.36)	4.29 (2.81)	31.0%	3.8%	20.1%
How distracting was Conflict Probe during the scenario?	5.86 (2.75)	2.38 (2.24)	2.31 (1.36)	2.24 (1.89)	5.8%	6.5%	4.7%
Were you prepared to trust the Conflict Probe during the scenario?	1.00 (0.00)	0.96 (0.14)	1.00 (0.00)	0.96 (0.13)	99.8%	99.7%	96.9%
How much confidence did you have in the detection accuracy of the Conflict Probe during the scenario?	6.20 (2.65)	6.27 (1.72)	7.49 (1.94)	6.02 (2.78)	38.7%	33.5%	23.1%
How much confidence did you have in your ability to detect conflicts in the scenario?	8.01 (0.94)	8.17 (1.10)	7.93 (1.46)	8.02 (1.47)	72.7%	71.9%	60.1%
How much confidence did you have that you and the Conflict Probe together could detect conflicts in the scenario?	8.29 (1.48)	8.44 (1.33)	8.37 (1.77)	8.27 (1.67)	64.7%	65.0%	51.6%
I responded to a conflict without question after I received an alert from the Conflict Probe.	2.68 (1.36)	1.96 (1.33)	3.45 (1.55)	2.10 (1.35)	51.2%	15.4%	41.2%
I noticed alerts from the Conflict Probe and believed that it was correct, but I checked other data before I responded to the conflicts.	3.09 (1.03)	2.12 (1.19)	3.36 (1.20)	2.63 (1.67)	58.2%	23.0%	54.4%
I noticed alerts from the Conflict Probe and did not know if they were correct, so I checked other data before responding to the conflicts.	3.74 (0.58)	2.33 (1.35)	3.18 (1.24)	2.59 (1.55)	70.9%	14.8%	36.2%
I detected conflicts on my own and did not know if I was correct, so I checked the alerts from the Conflict Probe before responding to the conflicts.	2.40 (1.30)	1.77 (0.94)	2.37 (1.29)	1.88 (0.99)	53.8%	21.5%	42.2%
I detected conflicts on my own and believed that I was correct, but I checked the alerts from the Conflict Probe before responding to the conflicts.	2.15 (0.96)	1.73 (0.90)	2.51 (1.25)	1.91 (0.96)	73.3%	49.3%	64.5%
I detected and resolved conflicts without using the Conflict Probe.	3.61 (0.66)	4.24 (0.82)	3.03 (1.05)	3.96 (1.11)	58.53%	19.7%	61.0%

Table J2. Summary of Bayesian Output for Radar Surveillance Noise by Location Analysis of the PSQ

Question	Mean *(SD)*				% within the ROPE		
	No/R & RA	No/RA	Yes/R & RA	Yes/RA	Noise	Location	Interaction
Rate the overall difficulty of this scenario.	5.74 (2.23)	5.98 (2.26)	5.37 (2.26)	5.83 (2.17)	50.1%	47.3%	43.9%
Rate your overall level of ATC performance.	7.56 (2.15)	7.36 (1.40)	7.64 (1.87)	7.60 (1.25)	61.1%	55.8%	48.3%
Rate your level of performance in resolving conflicts.	8.33 (1.20)	8.05 (1.51)	8.16 (1.83)	8.30 (1.07)	68.2%	68.1%	49.2%
Rate your level of performance in all other tasks.	7.93 (1.04)	8.13 (1.03)	7.88 (0.75)	8.00 (0.86)	87.5%	85.0%	73.8%
Rate your overall workload.	7.27 (1.29)	6.93 (1.63)	7.29 (1.27)	7.15 (1.11)	72.4%	68.1%	58.4%
Rate your overall level of situation awareness.	8.02 (1.57)	7.90 (1.51)	7.63 (2.06)	8.15 (1.24)	63.7%	60.7%	41.8%
Rate the performance of the simulation pilots in terms of their responses to control instructions and providing callbacks.	8.21 (2.52)	8.50 (1.72)	8.51 (1.77)	7.32 (1.99)	44.5%	43.7%	21.6%
What was the average number of aircraft in the sector?	15.79 (3.86)	16.12 (3.25)	15.88 (4.08)	16.48 (3.20)	33.8%	31.8%	29.9%
What was the largest number of aircraft in the sector at one time?	21.79 (4.82)	21.25 (2.64)	20.88 (3.94)	21.80 (3.90)	31.6%	30.8%	23.1%
Of all actual/probable conflicts, what percentage were never alerted by Conflict Probe?	3.97 (5.63)	3.94 (15.33)	4.84 (6.55)	4.15 (7.80)	12.4%	10.6%	10.9%
Of all Conflict Probe alerts, what percentage were false alerts?	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	100.0%	100.0%	100.0%
What was the total number of alerts?	14.89 (10.48)	13.71 (11.42)	15.08 (8.62)	14.64 (13.59)	10.3%	9.7%	10.7%
Of all accurate Conflict Probe alerts, what percentage would you describe as nuisance alerts, that is, accurate but of no use to you and you would prefer not to have them?	0.00 (0.01)	1.31 (2.95)	0.00 (0.00)	0.00 (0.02)	55.4%	55.4%	38.8%
When Conflict Probe gave accurate alerts, what was the average warning time before the conflict would have occurred?	444.41 (214.71)	415.22 (277.04)	454.22 (213.93)	454.57 (227.78)	0.5%	0.4%	0.6%
How accurate was Conflict Probe during the scenario?	7.62 (1.73)	6.34 (2.85)	7.49 (2.01)	6.04 (2.99)	43.8%	17.4%	36.0%
How useful was Conflict Probe during the scenario?	8.33 (1.63)	3.50 (2.85)	7.95 (2.18)	3.91 (2.60)	44.2%	0.0%	29.3%
How distracting was Conflict Probe during the scenario?	3.12 (1.78)	3.36 (2.91)	2.27 (1.33)	2.39 (1.82)	27.1%	49.0%	40.9%
Were you prepared to trust the Conflict Probe during the scenario?	1.00 (0.00)	0.97 (0.13)	1.00 (0.00)	0.97 (0.13)	100%	100%	99.1%
How much confidence did you have in the detection accuracy of the Conflict Probe during the scenario?	7.92 (1.51)	6.28 (2.47)	7.68 (1.86)	5.89 (2.71)	46.5%	8.0%	38.2%
How much confidence did you have in your ability to detect conflicts in the scenario?	8.35 (1.38)	7.48 (1.12)	8.01 (1.55)	7.87 (1.59)	67.6%	46.9%	39.6%
How much confidence did you have that you and the Conflict Probe together could detect conflicts in the scenario?	8.90 (1.25)	7.67 (1.12)	8.52 (1.69)	8.07 (1.61)	66.3%	24.9%	36.8%
I responded to a conflict without question after I received an alert from the Conflict Probe.	3.35 (1.22)	1.75 (1.15)	3.59 (1.49)	2.01 (1.32)	66.0%	1.7%	54.0%
I noticed alerts from the Conflict Probe and believed that it was correct, but I checked other data before I responded to the conflicts.	3.35 (0.96)	2.12 (1.79)	3.36 (1.25)	2.35 (1.89)	68.3%	13.1%	51.6%
I noticed alerts from the Conflict Probe and did not know if they were correct, so I checked other data before responding to the conflicts.	2.41 (0.94)	2.19 (1.59)	3.02 (1.29)	2.71 (1.60)	44.1%	63.6%	56.3%
I detected conflicts on my own and did not know if I was correct, so I checked the alerts from the Conflict Probe before responding to the conflicts.	2.01 (0.87)	1.73 (0.89)	2.28 (1.22)	1.89 (0.95)	78.5%	68.0%	68.5%
I detected conflicts on my own and believed that I was correct, but I checked the alerts from the Conflict Probe before responding to the conflicts.	2.38 (1.19)	1.85 (1.07)	2.52 (1.31)	1.80 (1.04)	79.9%	37.7%	60.8%
I detected and resolved conflicts without using the Conflict Probe.	3.27 (1.16)	4.34 (0.89)	2.97 (1.11)	4.01 (1.18)	68.0%	8.3%	64.7%

Table J3. Summary of Bayesian Output for Algorithm by Location Analysis of the OTS Ratings

Question	Mean *(SD)*				% within the ROPE		
	Legacy/ R & RA	Legacy/ RA	Enhancements/ R + RA	Enhancements/ RA	Algorithm	Location	Interaction
Maintaining Safe and Efficient Traffic Flow:							
Maintaining Separation and Resolving Potential Conflicts	7.71 (0.52)	7.77 (0.55)	7.53 (0.75)	7.65 (0.66)	91.5%	93.6%	83.0%
Sequencing Arrival and Departure Aircraft Efficiently	7.91 (0.21)	7.99 (0.00)	7.93 (0.15)	7.90 (0.21)	98.9%	99.2%	89.5%
Using Control Instructions Effectively	8.00 (0.06)	8.00 (0.00)	8.00 (0.00)	7.98 (0.07)	100.0%	100.0%	99.3%
Overall Safe and Efficient Traffic Flow	7.73 (0.47)	7.81 (0.44)	7.46 (0.77)	7.69 (0.59)	87.9%	90.8%	77.8%
Maintaining Attention and Situation Awareness:							
Maintaining Awareness of Aircraft Positions	7.42 (0.54)	7.93 (0.16)	7.00 (0.75)	7.81 (0.37)	80.8%	28.7%	65.2%
Ensuring Positive Control	7.64 (0.57)	7.78 (0.48)	7.54 (0.71)	7.73 (0.47)	95.3%	90.1%	83.9%
Detecting Pilot Deviations from Control Instructions	7.39 (0.91)	7.70 (0.58)	7.42 (0.63)	7.64 (0.60)	95.7%	82.5%	81.2%
Correcting Own Errors in a Timely Manner	7.56 (0.76)	7.91 (0.21)	7.50 (0.69)	7.68 (0.73)	88.7%	79.3%	73.1%
Overall Attention and Situation Awareness	7.41 (0.71)	7.83 (0.36)	7.09 (0.46)	7.75 (0.43)	88.2%	43.1%	71.3%
Prioritizing:							
Taking Actions in an Appropriate Order of Importance	7.82 (0.30)	7.92 (0.16)	7.37 (0.71)	7.82 (0.30)	85.1%	84.7%	64.2%
Preplanning Control Actions	8.00 (0.00)	8.00 (0.00)	7.25 (0.46)	7.97 (0.12)	77.9%	83.5%	23.0%
Handling Control Tasks for Several Aircraft	8.00 (0.00)	8.00 (0.00)	7.32 (0.52)	7.87 (0.21)	69.9%	87.5%	43.0%
Overall Prioritizing	7.75 (0.40)	7.87 (0.27)	7.33 (0.75)	7.75 (0.39)	83.8%	83.5%	67.9%
Providing Control Information:							
Providing Essential ATC Information	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Providing Additional ATC Information	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Overall Providing Control Information	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Technical Knowledge:							
Showing Knowledge of LOAs and SOPs	8.00 (0.00)	8.00 (0.00)	7.96 (0.12)	7.99 (0.13)	99.8%	99.9%	98.4%
Showing Knowledge of Aircraft Capabilities and Limitations	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Technical Knowledge: Overall Technical Knowledge	8.00 (0.00)	8.00 (0.00)	7.99 (0.05)	8.00 (0.00)	100.0%	100.0%	99.8%
Voice Communications:							
Using Proper Phraseology	8.00 (0.00)	8.00 (0.00)	7.99 (0.05)	8.00 (0.00)	100.0%	100.0%	100.0%
Communicating Clearly and Efficiently	8.00 (0.00)	8.00 (0.00)	7.99 (0.05)	8.00 (0.00)	100.0%	100.0%	100.0%
Listening for Pilot Readbacks and Requests	8.00 (0.00)	8.00 (0.00)	7.99 (0.05)	8.00 (0.00)	100.0%	100.0%	100.0%
Overall Voice Communication	8.00 (0.00)	8.00 (0.00)	7.99 (0.05)	8.00 (0.00)	100.0%	100.0%	100.0%
Conflict Probes:							
Using Conflict Probes in Appropriate Situations	7.83 (0.46)	6.57 (1.48)	7.32 (0.92)	5.16 (1.46)	20.2%	1.9%	32.4%
Taking Full Advantage of Conflict Probes in a Timely and Efficient Manner	7.40 (1.49)	6.28 (2.03)	7.20 (1.42)	4.81 (2.02)	29.5%	5.0%	24.5%
Conflict Probes: Overall Conflict Probe Use	7.87 (0.33)	6.55 (1.64)	7.70 (0.59)	5.05 (1.47)	28.6%	1.1%	21.5%

Table J4. Summary of Bayesian Output for Radar Surveillance Noise by Location Analysis of the OTS Ratings

Question	Mean *(SD)*				% within the ROPE		
	No/ R & RA	No/ RA	Yes/ R & RA	Yes/ RA	Algorithm	Location	Interaction
Maintaining Safe and Efficient Traffic Flow:							
Maintaining Separation and Resolving Potential Conflicts	7.45 (0.65)	7.20 (0.73)	7.43 (0.82)	7.46 (0.83)	91.6%	92.0%	69.5%
Sequencing Arrival and Departure Aircraft Efficiently	7.39 (1.13)	6.92 (1.25)	7.75 (0.60)	7.65 (0.68)	46.3%	70.6%	56.5%
Using Control Instructions Effectively	7.98 (0.10)	8.00 (0.00)	8.00 (0.00)	7.96 (0.12)	99.9%	99.9%	96.8%
Overall Safe and Efficient Traffic Flow	7.45 (0.66)	7.23 (0.92)	7.36 (0.86)	7.46 (0.84)	90.8%	90.7%	65.0%
Maintaining Attention and Situation Awareness:							
Maintaining Awareness of Aircraft Positions	7.45 (0.69)	7.24 (1.43)	6.98 (0.94)	7.47 (0.86)	81.4%	80.1%	40.8%
Ensuring Positive Control	7.35 (0.84)	7.23 (0.92)	7.44 (0.84)	7.58 (0.64)	83.9%	92.7%	68.5%
Detecting Pilot Deviations from Control Instructions	7.59 (0.60)	7.46 (0.77)	7.47 (0.63)	7.59 (0.60)	96.9%	96.8%	75.1%
Correcting Own Errors in a Timely Manner	7.49 (0.69)	7.39 (1.11)	7.40 (0.85)	7.36 (1.11)	87.2%	85.5%	72.2%
Overall Attention and Situation Awareness	7.45 (0.65)	7.31 (1.05)	7.20 (0.72)	7.58 (0.63)	93.3%	90.8%	51.2%
Prioritizing:							
Taking Actions in an Appropriate Order of Importance	7.34 (0.85)	7.34 (0.97)	7.23 (0.93)	7.58 (0.65)	90.5%	85.7%	62.8%
Preplanning Control Actions	7.46 (1.03)	7.37 (1.19)	7.30 (0.83)	7.46 (1.03)	83.7%	84.1%	63.0%
Handling Control Tasks for Several Aircraft	7.32 (1.00)	7.54 (0.82)	7.30 (0.79)	7.61 (0.64)	92.4%	79.7%	76.5%
Overall Prioritizing	7.28 (1.03)	7.43 (0.83)	7.21 (0.92)	7.59 (0.65)	92.0%	78.0%	70.3%
Providing Control Information:							
Providing Essential ATC Information	7.99 (0.05)	7.97 (0.19)	8.00 (0.00)	7.99 (0.05)	99.7%	99.6%	97.6%
Providing Additional ATC Information	7.98 (0.08)	7.93 (0.27)	8.00 (0.02)	7.98 (0.08)	98.9%	98.5%	94.6%
Overall Providing Control Information	8.00 (0.00)	7.95 (0.14)	8.00 (0.00)	7.99 (0.04)	99.3%	99.0%	95.7%
Technical Knowledge:							
Showing Knowledge of LOAs and SOPs	8.00 (0.00)	7.76 (0.63)	7.96 (0.15)	7.95 (0.20)	94.9%	96.9%	82.0%
Showing Knowledge of Aircraft Capabilities and Limitations	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Technical Knowledge: Overall Technical Knowledge	8.00 (0.00)	7.98 (0.14)	7.95 (0.14)	8.00 (0.00)	100.0%	100.0%	96.7%
Voice Communications:							
Using Proper Phraseology	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Communicating Clearly and Efficiently	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Listening for Pilot Readbacks and Requests	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Overall Voice Communication	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	100.0%	100.0%	100.0%
Conflict Probes:							
Using Conflict Probes in Appropriate Situations	7.85 (0.34)	4.79 (2.17)	7.50 (0.84)	4.96 (1.32)	62.6%	0.3%	43.3%
Taking Full Advantage of Conflict Probes in a Timely and Efficient Manner	7.97 (0.06)	3.80 (1.89)	7.89 (0.38)	4.46 (1.29)	56.2%	0.0%	29.1%
Conflict Probes: Overall Conflict Probe Use	7.98 (0.03)	4.69 (2.03)	7.90 (0.30)	4.87 (1.25)	58.4%	0.1%	34.6%

Table J5. Summary of Bayesian Output for Algorithm by Location Analysis of Simulation Pilot Commands

Command	Mean *(SD)*				% within the ROPE		
	Legacy/ R & RA	Legacy/ RA	Enhancements/ R & RA	Enhancements/ RA	Algorithm	Location	Interaction
Altitude	21.01 (4.62)	20.20 (5.69)	18.16 (4.05)	17.56 (5.07)	7.1%	23.0%	21.8%
Direct	9.69 (4.82)	8.66 (4.18)	8.90 (4.08)	9.94 (2.39)	29.9%	30.6%	18.2%
Heading	1.31 (1.29)	0.08 (0.18)	1.04 (1.12)	0.78 (1.20)	90.11%	55.8%	47.8%
Route	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	100.0%	100.0%	100.0%
Speed	0.90 (1.10)	1.17 (1.02)	1.18 (1.28)	1.21 (1.50)	75.2%	75.1%	60.3%

Table J6. Summary of Bayesian Output for Radar Surveillance Noise by Location Analysis of Simulation Pilot Commands

Command	Mean *(SD)*				% within the ROPE		
	No/ R & RA	No/ RA	Yes/ R & RA	Yes/ RA	Noise	Location	Interaction
Altitude	20.13 (5.25)	18.98 (3.34)	18.15 (3.93)	17.50 (4.87)	14.8%	23.2%	25.0%
Direct	9.37 (5.03)	8.87 (3.08)	8.89 (3.94)	9.95 (2.36)	31.8%	31.5%	21.7%
Heading	1.42 (2.04)	1.50 (2.13)	1.23 (1.61)	1.32 (1.72)	57.2%	59.3%	48.5%
Route	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	100.0%	100.0%	100.0%
Speed	0.61 (0.88)	0.39 (0.68)	1.21 (1.23)	1.10 (1.45)	35.8%	79.3%	64.5%

Appendix K: Means and Standard Deviations from Exit Questionnaire Items

Table K1. Summary of Means and (*SD*) to Compare Group Responses on Exit Questionnaire Items

Question	Group 1		Group 2		*t*-value	*p*-value
	Mean	*SD*	Mean	*SD*		
Rate the overall realism of the simulation.	7.00	0.82	7.00	2.16	0.00	1.00000
Rate the realism of the simulation hardware compared to actual equipment.	6.00	1.41	8.25	1.26	-2.38	0.05498
Rate the realism of the simulation software compared to actual equipment.	4.50	1.73	8.00	0.82	3.66	0.01064
Rate the realism of the simulation traffic scenarios compared to actual NAS traffic.	6.50	1.00	6.00	1.41	0.58	0.58470
To what extent did the WAK online workload rating technique interfere with your ATC performance?	4.25	2.21	4.00	2.31	0.16	0.88102
How effective was the training provided?	7.75	1.26	7.50	0.58	0.36	0.73036
To what extent did the fNIRS interfere with your ATC performance?	5.50	2.65	5.00	4.36	0.19	0.85635
Of all actual/probable conflicts in the field with standard, legacy settings, what percentage are never alerted by Conflict Probe?	10.00	7.07	8.75	2.50	0.33	0.75022
Of all Conflict Probe alerts in the field with standard, legacy settings, what percentage are false alerts?	10.00	7.07	16.25	4.79	1.46	0.19357
Of all accurate Conflict Probe alerts in the field with standard, legacy settings, what percentage would you describe as nuisance alerts?	10.75	6.99	16.25	4.79	1.30	0.24199
When Conflict Probe gives accurate alerts in the field with standard, legacy settings, what is the average warning time before the conflict would have occurred?	480.00	146.97	390.00	186.55	0.76	0.47719
How accurate is Conflict Probe?	7.75	0.50	6.75	0.50	2.83	0.03002
How useful is Conflict Probe when you are working in a team at the RA-Side?	6.00	1.63	5.75	1.89	0.20	0.84809
How useful is Conflict Probe when you are working in a team at the R-Side?	8.25	1.71	4.75	2.75	2.16	0.07405
How useful is Conflict Probe when you are working alone at the R-Side?	8.00	2.31	4.75	3.30	1.61	0.15799
If you were working in a team at the R-Side, how distracting would it be to have Conflict Probe displayed on the R-Side?	4.00	3.16	5.00	4.08	0.39	0.71190
If you were working alone at the R-Side, how distracting would it be to have Conflict Probe displayed on the R-Side?	3.50	2.38	5.00	4.08	0.63	0.54898

Table K2. Summary of Means and (*SD*) to Compare Group Ratings of CP Features in the Exit Questionnaire

	Group 1		Group 2			
Question: How did the following Conflict Probe features affect your performance in this study?	**Mean**	***SD***	**Mean**	***SD***	***t*-value**	***p*-value**
Conflict Probe alerts in the Conflict Alert List when they were correct.	6.75	2.36	7.00	1.63	0.17	0.86753
Conflict Probe alerts in aircraft data blocks when they were correct.	9.50	0.58	7.75	1.71	1.94	0.10024
The fact that an alert portal did not appear until there was a red alert for the R-Side.	5.50	2.65	7.00	1.41	1.00	0.35592
The fact that yellow alerts were only presented on the RA-Side, not on the R-Side.	5.25	3.77	6.50	1.29	0.63	0.55398

Appendix L: Data Tables for Analysis on Types of Clearances

Table L1. List of Expedited Clearances

ACID	Expedite to Altitude	Time	RA-Side Probe Alert Color	R-Side Probe Color	CA Time	Expedite Given Before CA	Proactive or Reactive
No Alert for Aircraft on R-Side							
LXJ313	E240	37:16.0	Yellow/Red		37:08.0	Yes**	Reactive
LXJ852	E240	35:49.0	Red				Proactive
ASQ4074	E240	39:11.0	Muted Red				Reactive
DAL337	E240	31:51.0	Red		32:24.0	Yes*	Reactive
LXJ206	E240	39:05.0	Muted Yellow				Proactive
ASQ9302	E330	20:32.0	Red				Reactive
ASQ6964	E280	31:37.0	Red		31:16.0	No	Reactive
ASQ7052	E350	25:12.0	Red		47:00.0	No	Reactive
DAL692	E290	30:29.0	Red		30:14.0	No	Reactive
EGF562	E310	26:07.0	Muted Yellow				Proactive
AAL869	E400	04:19.0	Yellow				Reactive
ASQ4921	E310	22:11.0	None				Proactive
ASQ3884	E310	27:42.0	None				Proactive
ASQ2438	E290	32:11.0	Red		21:23.0	No	Reactive
ASQ8572	E250	08:35.0	Yellow				Reactive
No Alert for Aircraft on R-Side							
LXJ951	E260	29:03.0	Red	Red			Proactive
ASQ5114	E270	31:02.0	Red	Red	31:04.0	Yes	Proactive
TRS2199	E350	14:52.0	Red	Red			Proactive
ASQ7816	E300	25:27.0	Red	Red	25:18.0	Yes**	Reactive
AAL869	E400	03:38.0	Red	Red			Proactive
ASQ6191	E270	22:32.0	Red	Red	22:55.0	Yes	Proactive
DAL755	E280	31:37.0	Red	Red			Reactive

* Expedited command put aircraft into conflict sooner.

** For these flights, the time shows the pilot command entered after the activation of the CA; however, the clearance was given prior and the simulation pilot was delayed entering it.

Table L2. List of Altitude Clearances Amended After Controller Issuance

ACID	Commanded Altitude	First Clearance Time	Second Clearance Time	Time Lapse (mm:ss)	RA-Side Probe Alert Color	R-Side Probe Alert Color	CA Time
No Alert for Aircraft on R-Side							
DAL337	E240	31:40	32:24	00:44	Red	None	31:22
ASQ9302	A350	20:40	21:18	00:38	Red	None	21:05
ASQ9302	A330	21:18	23:18	02:00	Red	None	22:55
SWA803	A240	14:47	15:28	00:41	Red	None	15:24
LXJ313	A280	33:05	33:44	00:39	Red	None	33:40
ASQ1129	A270	25:28	26:15	00:47	Red	None	26:06
DAL337	A290	31:37	32:03	00:26	Red	None	
DAL337	E240	31:40	32:23	00:43	Red	None	32:23
DAL755	A240	30:09	32:36	02:27	Yellow	None	32:22
ASQ5967	A330	28:48	29:10	00:22	Yellow	None	
ASQ7646	A270	27:04	29:04	02:00	Muted Red	None	
ASQ7018	A270	24:38	25:42	01:04	Red		25:17
			Average Time Lapse =	***01:03***			
Alert for Aircraft on R-Side							
ASQ3787	A310	20:58	21:03	00:05	Red	Red	
ASQ3787	A300	21:07	21:22	00:15	Red	Red	
DAL337	A240	26:38	27:18	00:40	Red	Red	27:09
ASQ5967	A270	25:07	26:03	00:56	Red	Red	47:00
ASQ5680	A270	30:55	31:58	01:03	Red	Red	31:36
ASQ7816	A270	24:37	24:47	00:10	Red	Red	18:00
ASQ5950	A280	29:20	29:45	00:25	Red	Red	
ASQ3787	A310	18:30	18:42	00:12	Red	Red	
ASQ3787	A290	18:42	18:53	00:11	Red	Red	
SWA665	A240*	12:16	12:35	00:19	Red	Red	
			Average Time Lapse =	***00:26***			

* Issued by controller but not entered by simulation pilot.

Table L3. List of Altitude Clearances Amended Before Controller Issuance

Command	Command Type	Command End Time	R-Side Probe Color
QQ 310 035	Interim Altitude	23:09.1	Red
QQ 300 035	Interim Altitude	23:13.9	
	Time between entries	00:04.8	
QQ 310 025	Interim Altitude	19:03.4	Red
QQ 300 025	Interim Altitude	19:06.6	
	Time between entries	00:03.2	
QZ 240 065	Assigned Altitude	32:28.3	Red
QQ 310 065	Interim Altitude	32:34.6	
	Time between entries	00:06.4	
QZ 240 005	Assigned Altitude	20:54.8	Red
QQ 250 005	Interim Altitude	21:00.3	
	Time between entries	00:05.6	
QQ 280 011	Interim Altitude	27:18.8	Yellow
QQ 260 011	Interim Altitude	27:23.2	
	Time between entries	00:04.4	
QZ 240 089	Assigned Altitude	10:48.1	Red
QQ 280 089	Interim Altitude	10:52.7	
	Time between entries	00:04.6	
QQ 310 024	Interim Altitude	16:56.9	Red
QQ 300 024	Interim Altitude	17:01.9	
	Time between entries	00:05.0	
QZ 280 107	Assigned Altitude	29:35.5	Red
QQ 260 107	Interim Altitude	29:40.4	
	Time between entries	00:04.9	
QZ 280 094	Assigned Altitude	14:04.3	Red
QQ 260 094	Interim Altitude	14:12.4	
	Time between entries	00:08.1	
QZ 240 005	Assigned Altitude	23:18.3	Red
QQ 270 005	Interim Altitude	23:26.4	
	Time between entries	00:08.2	
QQ 260 020	Interim Altitude	27:17.4	Red
QQ 250 020	Interim Altitude	27:26.5	
	Time between entries	00:09.1	
QZ 310 018	Assigned Altitude	19:52.0	Red
QZ 290 018	Assigned Altitude	20:00.0	
	Time between entries	00:08.0	
	Average Time between entries =	***00:06.0***	

Appendix M: Detailed PTT Results

Tables M1 and M2 show the mean and standard deviation of the PTT durations and counts for each of the six analyzed conditions. We included relevant subsets of these means in each of our analyses.

Table M1. Mean and (*SD*) Percentage of Time on Frequency for All Analyzed Conditions

Condition	Mean and (*SD*) of % Time on Frequency
R & RA/Enhancements/No Noise	24.31% (9.00%)
R & RA/Legacy/Noise	24.25% (8.01%)
R & RA/Enhancements/Noise	23.61% (8.63%)
R & RA Overall	24.06% (8.54%)
RA Only/Enhancements/No Noise	23.64% (8.61%)
RA Only/Legacy/Noise	23.32% (9.45%)
RA Only/Enhancements/Noise	23.24% (8.39%)
RA Only Overall	23.40% (8.81%)

Table M2. Mean and (*SD*) Transmissions per Hour for All Analyzed Conditions

Condition	Mean and (*SD*) of % Time on Frequency
R & RA/Enhancements/No Noise	246.46 (86.38)
R & RA/Legacy/Noise	244.66 (71.39)
R & RA/Enhancements/Noise	229.07 (80.93)
R & RA Overall	240.06 (80.00)
RA Only/Enhancements/No Noise	236.67 (80.23)
RA Only/Legacy/Noise	225.62 (82.31)
RA Only/Enhancements/Noise	233.86 (77.71)
RA Only Overall	232.05 (80.06)

The means and SDs for the subsets of conditions used in each analysis on PTT Duration and Count appear in Tables M3 through M8. Each table is followed by the results of the regression analysis for the effects on the given variable addressed by the given subset of conditions.

Table M3. Mean and (*SD*) Percentage of Time on Frequency, Overall by Location

Condition	Mean and (*SD*) of % Time on Frequency
R & RA Overall	24.06% (8.54%)
RA Only Overall	23.40% (8.81%)

As reported in Section 3.5.1, the effect of Location was significant, $F(1, 805) = 7.33$, $p < .0070$.

Table M4. Mean and (*SD*) Transmissions per Hour, Overall by Location

Condition	Mean and (*SD*) of % Time on Frequency
R & RA Overall	240.06 (80.00)
RA Only Overall	232.05 (80.06)

As reported in Section 3.5.1, the effect of Location was significant, $F(1, 805.1) = 6.82$, $p < .0093$.

Table M5. Mean and (*SD*) Percentage of Time on Frequency by Location and Algorithm

Condition	Mean and (*SD*) of % Time on Frequency
R & RA/Legacy/Noise	24.25% (8.01%)
R & RA/Enhancements/Noise	23.61% (8.63%)
RA Only/Legacy/Noise	23.32% (9.45%)
RA Only/Enhancements/Noise	23.24% (8.39%)

The main effect of Location was significant, $F(1, 532.8) = 9.15$, $p < .0027$. The main effect of Algorithm was not significant, $F(1, 533) = 1.69$, $p < .20$. The interaction of Location and Algorithm was also not significant, $F(1, 533) = 0.25$, $p < .62$.

Table M6. Mean and (*SD*) Transmissions per Hour, by Location and Algorithm

Condition	Mean and (*SD*) of % Time on Frequency
R & RA/Legacy/Noise	244.66 (71.39)
R & RA/Enhancements/Noise	229.07 (80.93)
RA Only/Legacy/Noise	225.62 (82.31)
RA Only/Enhancements/Noise	233.86 (77.71)

The main effect of Location was significant, $F(1, 535.7) = 6.35, p < .012$. The main effect of Algorithm was not significant, $F(1, 536.1) = 0.97, p < .33$. The interaction of Location and Algorithm was significant, $F(1, 536.1) = 4.83, p < .029$. A post-hoc Tukey test comparing the four means that we analyzed for the interaction revealed that the only two means that differed from each other were the two Legacy conditions.

Table M7. Mean and (*SD*) Percentage of Time on Frequency, by Location and Noise

Condition	Mean and (*SD*) of % Time on Frequency
R & RA/Enhancements/No Noise	24.31% (9.00%)
R & RA/Enhancements/Noise	23.61% (8.63%)
RA Only/Enhancements/No Noise	23.64% (8.61%)
RA Only/Enhancements/Noise	23.24% (8.39%)

The main effect of Location showed a trend toward significance, $F(1, 533.2) = 3.50, p < .062$. The main effect of Noise was significant, $F(1, 533.2) = 5.07, p < .025$. The interaction of Location and Noise was not significant, $F(1, 533.3) = 1.28, p < .26$.

Table M8. Mean and (*SD*) Transmissions per Hour, by Location and Noise

Condition	Mean and (*SD*) of % Time on Frequency
R & RA/Enhancements/No Noise	246.46 (86.38)
R & RA/Enhancements/Noise	229.07 (80.93)
RA Only/Enhancements/No Noise	236.67 (80.23)
RA Only/Enhancements/Noise	233.86 (77.71)

The main effect of Location was not significant, $F(1, 536.1) = 0.95, p < .33$. The main effect of Noise was significant, $F(1, 536.1) = 7.67, p < .0059$. The interaction of Location and Noise was not significant, $F(1, 536.1) = 0.12, p < .74$.

Appendix N: Detailed Results on Controller Commands

Tables N1, N2, and N3 show the means we compared for each effect or combination of effects, with their SDs and the result of each component of the associated regression model. We report these results for each of the four primary command types.

Table N1. Mean and (*SD*) Hourly Rates of Key Controller Command Types by CP Location

Command Type	R & RA Side		RA Side Only		Statistical Significance
Altitude	44.01	(36.01)	42.08	(34.57)	$F(1, 726.2) = 0.14$, $p < .71$
Heading	4.86	(12.43)	2.73	(9.01)	$F(1, 550.1) = 6.27$, $p < .013$
Speed	2.03	(8.93)	2.22	(8.35)	$F(1, 459.6) = 0.34$, $p < .56$
Route	42.57	(52.64)	39.92	(46.53)	$F(1, 726.1) = 0.61$, $p < .81$

Table N2. Mean and (*SD*) Hourly Rates of Key Controller Command Types by CP Location and Algorithm

Command Type	R & RA, Legacy	R & RA, Enhancement	RA Only, Legacy	RA Only, Enhancement	Statistical Significance
Altitude	44.77 (33.25)	43.73 (34.56)	43.77 (36.82)	37.44 (30.99)	Location: $F(1, 472.8) = 0.61$, $p < .44$ Algorithm $F(1, 472.3) = 1.89$, $p < .17$ Loc. x Algo.: $F(1, 472.5) = 1.16$, $p < .29$
Heading	5.82 (13.39)	4.83 (11.99)	0.65 (4.38)	3.67 (9.88)	Location: $F(1, 357.2) = 9.19$, $p < .0027$ Algorithm $F(1, 357.1) = 0.80$, $p < .38$ Loc. x Algo.: $F(1, 357.1) = 3.98$, $p < .047$
Speed	1.10 (5.67)	3.33 (11.87)	3.08 (10.36)	2.80 (8.79)	Location: $F(1, 298.9) = 1.19$, $p < .28$ Algorithm $F(1, 297.3) = 1.12$, $p < .30$ Loc. x Algo.: $F(1, 297.5) = 1.49$, $p < .23$
Route	37.97 (48.58)	38.90 (45.12)	42.30 (45.27)	42.05 (49.05)	Location: $F(1, 472.4) = 2.24$, $p < .14$ Algorithm $F(1, 472.1) = 0.086$, $p < .77$ Loc. x Algo.: $F(1, 472.2) = 0.045$, $p < .84$

Table N3. Mean and (*SD*) Hourly Rates of Key Controller Command Types by CP Location and Noise

Command Type	R & RA, No Noise	R & RA, Noise	RA Only, No Noise	RA Only, Noise	Statistical Significance
Altitude	43.51 (39.94)	43.73 (34.56)	44.88 (35.31)	37.44 (30.99)	Location: $F(1, 474.3) = 0.36$, $p < .55$ Noise: $F(1, 474.3) = 1.08$, $p < .31$ Loc. x Noise: $F(1, 474.7) = 1.44$, $p < .24$
Heading	3.94 (11.85)	4.83 (11.99)	3.91 (11.09)	3.67 (9.88)	Location: $F(1, 357) = 0.36$, $p < .55$ Noise: $F(1, 357.1) = 0.18$, $p < .68$ Loc. x Noise: $F(1, 357.2) = 0.19$, $p < .67$
Speed	1.81 (8.57)	3.33 (11.87)	0.78 (4.80)	2.80 (8.79)	Location: $F(1, 297.4) = 0.25$, $p < .62$ Noise: $F(1, 297.6) = 3.79$, $p < .053$ Loc. x Noise: $F(1, 298.6) = 0.35$, $p < .56$
Route	50.38 (61.51)	38.90 (45.12)	35.45 (45.33)	42.05 (49.05)	Location: $F(1, 474.1) = 1.13$, $p < .29$ Noise: $F(1, 474.1) = 0.14$, $p < .72$ Loc. x Noise: $F(1, 474.3) = 6.21$, $p < .014$

Tables N4 through N6 present the mean number of controller commands per hour of each individual type as computed by our data reduction process, by each of the condition sets we compared in the statistical analyses discussed above. We use the total valid (accepted by the system, excluding typographical errors and invalid inputs) commands of each type entered from the R-Side position during the core times, converted to hourly rates. We include all eight participants in these means, even those who did not use certain types of commands. Note that for the statistical analyses presented in Section 3.5.2 and in the preceding tables, the total for Altitude commands was the sum of Acknowledge IC Altitude, Assigned Altitude, Interim Altitude, and Remove Interim Altitude; the Heading command total was the sum of Heading and Remove Heading; Speed was the sum of Remove Speed and Speed; and Route was the sum of Route, Route Display, and Route Hide.

Table N4. Mean Hourly Rates of All Controller Command Types by CP Location

Command Type	R & RA Side	RA Side Only
Assigned Altitude	21.07	20.79
Drop Track	0.07	0.07
Flight Plan Readout	1.00	0.43
Halo	11.57	9.07
Handoff	6.86	4.71
HandoffAccept	96.29	94.57
Heading	2.07	1.07
Inhibit AutoHandoff	0.00	0.07
Interim Altitude	14.43	11.07
Leader Direction	210.57	211.43
Leader Direction/Length	16.29	20.43
Leader Length	24.79	27.21
Macro	1.64	0.86
PointOut	0.86	0.64
PVD Select	0.64	0.36
Quick Look	0.07	0.00
Range/Bearing	1.14	1.21
Remove Heading	1.57	0.71
Remove Interim Altitude	3.79	4.14
Remove Speed	0.07	0.00
Route	18.36	18.00
Route Display	8.50	7.29
Route Hide	12.93	10.43
Select	100.00	93.86
Speed	1.14	1.21
Start Track	1.79	1.29
Suppress Conflict Alert	0.43	0.29

Table N5. Mean Hourly Rates of All Controller Command Types by Location and Algorithm

Command Type	R & RA Legacy	R & RA Enhancement	RA Only Legacy	RA Only Enhancement
Assigned Altitude	23.36	19.50	21.64	19.93
Drop Track	0.00	0.21	0.00	0.00
Flight Plan Readout	1.71	0.43	0.21	0.64
Halo	10.71	8.79	13.07	6.00
Handoff	7.50	6.64	4.50	4.50
HandoffAccept	96.64	95.14	95.79	92.79
Heading	2.57	1.71	0.21	1.29
Interim Altitude	15.00	12.86	13.07	9.64
Leader Direction	210.21	201.21	205.50	208.71
Leader Direction/Length	13.93	15.43	20.14	18.21
Leader Length	20.36	27.43	32.79	23.36
Macro	1.93	1.50	1.50	1.07
PointOut	0.21	0.86	0.86	0.21
PVD Select	0.43	0.64	0.21	0.21
Quick Look	0.00	0.21	0.00	0.00
Range/Bearing	1.71	1.29	0.86	1.71
Remove Heading	1.50	1.93	0.21	1.07
Remove Interim Altitude	3.43	3.64	3.86	1.71
Remove Speed	0.00	0.21	0.00	0.00
Route	19.93	15.86	15.43	20.79
Route Display	4.07	9.00	9.86	7.07
Route Hide	10.93	10.50	12.43	9.21
Select	97.50	100.93	91.07	102.64
Speed	0.64	1.71	1.71	1.50
Start Track	0.86	3.64	0.86	1.93
Suppress Conflict Alert	0.43	0.00	0.00	0.86

Note. This table includes only the conditions used in analyzing the Location and Algorithm effects. Surveillance Noise was present in all conditions.

Table N6. Mean Hourly Rates of All Controller Command Types by Location and Noise

Command Type	R & RA No Noise	R & RA Noise	RA Only No Noise	RA Only Noise
Assigned Altitude	20.36	19.50	20.79	19.93
Drop Track	0.00	0.21	0.21	0.00
Flight Plan Readout	0.86	0.43	0.43	0.64
Halo	15.21	8.79	8.14	6.00
Handoff	6.43	6.64	5.14	4.50
HandoffAccept	97.07	95.14	95.14	92.79
Heading	1.93	1.71	1.71	1.29
Inhibit AutoHandoff	0.00	0.00	0.21	0.00
Interim Altitude	15.43	12.86	10.50	9.64
Leader Direction	220.29	201.21	220.07	208.71
Leader Direction/Length	19.50	15.43	22.93	18.21
Leader Length	26.57	27.43	25.50	23.36
Macro	1.50	1.50	0.00	1.07
PointOut	1.50	0.86	0.86	0.21
PVD Select	0.86	0.64	0.64	0.21
Quick Look	0.00	0.21	0.00	0.00
Range/Bearing	0.43	1.29	1.07	1.71
Remove Heading	1.29	1.93	0.86	1.07
Remove Interim Altitude	4.29	3.64	6.86	1.71
Remove Speed	0.00	0.21	0.00	0.00
Route	19.29	15.86	17.79	20.79
Route Display	12.43	9.00	4.93	7.07
Route Hide	17.36	10.50	9.64	9.21
Select	101.57	100.93	87.86	102.64
Speed	1.07	1.71	0.43	1.50
Start Track	0.86	3.64	1.07	1.93
Suppress Conflict Alert	0.86	0.00	0.00	0.86

Note. This table includes only the conditions used in analyzing the Location and Noise effects. Algorithm was Algorithmic Enhancements in all conditions. Certain cells duplicate cells from the previous table because we used the same condition means in the analyses.

Appendix O: Detailed Results on Time/Distance and Aircraft Handled

Tables O1 through O3 present the mean and SD time (minutes) and distance (NM) for each of the three definitions of workload as a function of each of the manipulations we tested. As described in Section 3.6.1, we conducted significance testing only for the Time (Duration) metrics.

Table O1. Mean and (*SD*) Time and Distance in Sector by CP Location, and Statistical Significance for Time Metrics

Metric	R & RA Side		RA Side Only		Statistical Significance
Geographic Bounds - Duration	10.11	(3.95)	10.00	(3.99)	$F(1, 1346) = 0.0007$, $p < .98$
Geographic Bounds - Distance	74.09	(28.23)	73.13	(28.40)	
Under Control - Duration	11.31	(3.87)	11.31	(3.89)	$F(1, 1345) = 2.71$, $p < .11$
Under Control - Distance	82.90	(27.26)	82.66	(27.40)	
Under Responsibility - Duration	12.66	(3.92)	12.66	(3.96)	$F(1, 1346) = 3.42$, $p < .065$
Under Responsibility - Distance	92.74	(27.68)	92.55	(27.90)	

Table O2. Mean and (*SD*) Time and Distance in Sector by CP Location and Algorithm, and Statistical Significance for Time Metrics

Metric	R & RA, Legacy	R & RA, Enhancement	RA Only, Legacy	RA Only, Enhancement	Statistical Significance
Geographic Bounds - Duration	9.68 (3.86)	10.38 (4.01)	10.04 (3.90)	9.78 (4.10)	Location: $F(1, 770.3) = 0.24$, $p < .63$ Algorithm $F(1, 770) = 0.40$, $p < .53$ Loc. x Algo.: $F(1, 769.8) = 3.04$, $p < .084$
Geographic Bounds - Distance	71.17 (28.08)	75.86 (28.37)	73.32 (27.50)	71.48 (29.29)	
Under Control - Duration	10.92 (3.84)	11.53 (3.97)	11.41 (3.87)	11.10 (3.97)	Location: $F(1, 771.5) = 5.14$, $p < .024$ Algorithm $F(1, 770.8) = 0.17$, $p < .90$ Loc. x Algo.: $F(1, 770.3) = 1.25$, $p < .27$
Under Control - Distance	80.31 (27.76)	84.23 (27.57)	83.24 (26.99)	81.04 (28.03)	
Under Responsibility - Duration	12.23 (3.89)	12.92 (4.00)	12.78 (3.90)	12.45 (4.03)	Location: $F(1, 771) = 7.47$, $p < .0065$ Algorithm $F(1, 770.5) = 0.23$, $p <0.64$ Loc. x Algo.: $F(1, 770) = 2.83$, $p < .094$
Under Responsibility - Distance	89.89 (28.17)	94.39 (27.74)	93.31 (27.19)	90.94 (28.51)	

Table O3. Mean and (*SD*) Time and Distance in Sector by CP Location and Noise, and Statistical Significance for Time Metrics

Metric	R & RA, No Noise	R & RA, Noise	RA Only, No Noise	RA Only, Noise	Statistical Significance
Geographic Bounds - Duration	10.25 (3.94)	10.38 (4.01)	10.17 (3.96)	9.78 (4.10)	Location: $F(1, 800.4) = 1.11$, $p < .30$ Noise: $F(1, 800.6) = 0.15$, $p < .70$ Loc. x Noise: $F(1, 800.5) = 1.41$, $p < .24$
Geographic Bounds - Distance	75.15 (28.12)	75.86 (28.37)	74.51 (28.39)	71.48 (29.29)	
Under Control - Duration	11.48 (3.77)	11.53 (3.97)	11.41 (3.84)	11.10 (3.97)	Location: $F(1, 800.4) = 0.097$, $p < .76$ Noise: $F(1, 801.1) = 0.14$, $p < .72$ Loc. x Noise: $F(1, 800.8) = 0.0041$, $p < .95$
Under Control - Distance	84.06 (26.38)	84.23 (27.57)	83.64 (27.20)	81.04 (28.03)	
Under Responsibility - Duration	12.81 (3.86)	12.92 (4.00)	12.74 (3.95)	12.45 (4.03)	Location: $F(1, 800.5) = 0.031$, $p < .87$ Noise: $F(1, 801.1) = 0.062$, $p < .81$ Loc. x Noise: $F(1, 800.9) = 0.0002$, $p < 1.00$
Under Responsibility - Distance	93.86 (27.04)	94.39 (27.74)	93.35 (28.03)	90.94 (28.51)	

Table O-4 shows the mean and standard deviation of the number of aircraft handled per hours for each of the six analyzed conditions, along with the overall means by Location. We included relevant subsets of these means in each of our analyses.

Table O4. Mean and (*SD*) Aircraft Handled per Hour, for All Analyzed Conditions

Condition	Mean and (*SD*) Aircraft Handled per hour
R & RA/Enhancements/No Noise	66.00 (1.59)
R & RA/Legacy/Noise	62.57 (2.75)
R & RA/Enhancements/Noise	63.86 (1.77)
R & RA Overall	64.14 (2.47)
RA Only/Enhancements/No Noise	65.14 (2.90)
RA Only/Legacy/Noise	60.86 (1.59)
RA Only/Enhancements/Noise	63.21 (2.32)
RA Only Overall	63.07 (2.86)

The overall means for the three R & RA-Side conditions, vs. the three RA Only conditions, and for the subsets of conditions used in each analysis on WAK Ratings, appear in Tables O-5 through O-7. Each table is followed by the results of the regression analysis for the effects addressed by the subset of conditions.

Table O5. Mean and (*SD*) Aircraft Handled per Hour, Overall by Location

Condition	Mean and (*SD*) of Rating
R & RA Overall	64.14 (2.47)
RA Only Overall	63.07 (2.86)

The effect of Location was not significant, $F(1, 39) = 1.83$, $p < .19$.

Table O6. Mean and (*SD*) Aircraft Handled per Hour, by Location and Algorithm

Condition	Mean and (*SD*) of Rating
R & RA/Legacy/Noise	62.57 (2.75)
R & RA/Enhancements/Noise	63.86 (1.77)
RA Only/Legacy/Noise	60.86 (1.59)
RA Only/Enhancements/Noise	63.21 (2.32)

The main effect of Location was not significant, $F(1, 21) = 2.59$, $p < .137$. The main effect of Algorithm was significant, $F(1, 21) = 6.19$, $p < .022$. The interaction of Location and Algorithm was not significant, $F(1, 21) = 0.54$, $p < .48$.

Table O7. Mean and (*SD*) Aircraft Handled per Hour, by Location and Noise

Condition	Mean and (*SD*) of Rating
R & RA/Enhancements/No Noise	66.00 (1.59)
R & RA/Enhancements/Noise	63.86 (1.77)
RA Only/Enhancements/No Noise	65.14 (2.90)
RA Only/Enhancements/Noise	63.21 (2.32)

The main effect of Location was not significant, $F(1, 21) = 0.83$, $p < .38$. The main effect of Noise was significant, $F(1, 21) = 6.09$, $p < .023$. The interaction of Location and Noise was not significant, $F(1, 21) = 0.017$, $p < .90$.