NASA/TM–20220005742

The Effects of Training and Flight Director Use on Pilot Monitoring Performance: A Sensemaking Approach

Dorrit Billman
NASA Ames Research Center

Randall J. Mumaw
San Jose State University Foundation

Peter M. T. Zaal
Metis Technology Solutions, Inc

Thomas J. Lombaerts
KBR Wyle Services, LLC

Isabel Torron
San Jose State University Foundation

Saad Jamal
University of California, Berkeley

Megan Shyr
NASA Ames Research Center

Michael Feary
NASA Ames Research Center

December 2021

NASA STI Program…in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, and organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question via to help@sti.nasa.gov

- Phone the NASA STI Help Desk at (757) 864-9658

- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM—20220005742

The Effects of Training and Flight Director Use on Pilot Monitoring Performance: A Sensemaking Approach

Dorrit Billman
NASA Ames Research Center

Randall J. Mumaw
San Jose State University Foundation

Peter M. T. Zaal
Metis Technology Solutions, Inc

Thomas J. Lombaerts
KBR Wyle Services, LLC

Isabel Torron
San Jose State University Foundation

Saad Jamal
University of California, Berkeley

Megan Shyr
NASA Ames Research Center

Michael Feary
NASA Ames Research Center

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California

December 2021

Available from:

NASA STI Program
STI Support Services
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

This report is also available in electronic form at http://www.sti.nasa.gov
or http://ntrs.nasa.gov/

Table of Contents

List of Figures and Tables

Acronyms and Definitions

ADI .....attitude indicator
AGL .....above ground level
ANL .....AoI Neglect Latency
ANOVA.....analysis of variance
AoI .....area of interest
ATC .....air traffic control
CP .....confederate pilot
deg/s.....degrees per second
DM.....data manager
EFIS .....electronic flight instrument system
EICAS.....engine-indication and crew-alerting system
FAA .....Federal Aviation Administration
FD .....flight director
FMS .....flight management system
FO .....first officer
FPM .....flight path management
FSS.....Full Flight Simulator
FTD.....Flight Training Device
Hz.....hertz (a unit of frequency equal to one cycle per second)
ID .....identification
IP.....instructor pilot
MCP.....mode control panel
mil.....military
NASA .....National Aeronautics and Space Administration
NAV.....navigation
ND.....navigation display
NTP.....network time protocol
Obs.....observer
OTW .....out-the-window
PIC .....pilot in command
PDT.....Proportion Dwell Time
PF .....pilot flying
PFD .....primary flight display
PIC .....pilot in command
PM.....pilot monitoring
SD .....standard deviation
SM.....simulator manager
T/D.....top-of-descent
UTC .....universal coordinated time
VSD .....vertical situation display

The Effects of Training and Flight Director Use on Pilot Monitoring Performance: A Sensemaking Approach

Dorrit Billman, Randall J. Mumaw, Peter M.T. Zaal,
Thomas J. Lombaerts, Isabel Torron, Saad Jamal,
Megan Shyr, and Michael Feary

Abstract

The need for improved pilot monitoring and awareness has been widely recognized, and training is a possible intervention. Based on our sensemaking-model of monitoring, we identified key properties of monitoring flight path. We designed scenarios with associated behavioral markers that provide measures of monitoring performance and a short training module emphasizing our proactive, anticipatory view of monitoring. Nineteen first officers from a major U.S. airline participated in the training study. Each pilot flew in a simulator pretraining session, participated in a training session, and flew in a simulator post-training session. We found modest but significant improvements in monitoring. The study collected video, simulator, and eye-tracking data and also manipulated whether the Flight Director was on or off. Limitations and future directions are discussed.

1. Introduction

The commercial aviation industry world-wide has identified a need for improved pilot monitoring and awareness (e.g., FAA, 2013; ICAO, 2016). More specifically, aviation safety data indicate that failures in pilots' monitoring and awareness relative to flight path management (FPM) have contributed to a range of undesired outcomes: accidents, major upsets, and non-compliance with air traffic control (ATC) guidance. The Federal Aviation Administration (FAA) has further stated that these types of FPM failures are likely to worsen with the increasingly complex air traffic control systems and FPM concepts proposed for NextGen (https://www.faa.gov/nextgen/this_is_nextgen/today/) operations. Adding to this complexity is the introduction of increasingly automated aircraft systems that can increase monitoring burdens.

One potential mitigation for this situation is to enhance pilot training for monitoring. A recent exploration of monitoring skill (Billman, et al., 2020; Mumaw, et al., 2020) has characterized monitoring as it relates to FPM, identified component skills and knowledge required for effective monitoring, and then reviewed the literature on the effectiveness of various training approaches for improving monitoring and awareness. One recommendation from these reports was that monitoring is best cast as a "sensemaking" activity which builds up the pilots' understanding of the dynamic

situation. Effective monitoring is an active, strategic process of "flying ahead of the airplane." Specifically, the pilot engages in continuous cycles of three activities (see Figure 1):

- Identifying what is most important to update or assess in their understanding of the situation: the current or upcoming state of the airplane systems, of the operational environment, or of the airplane's position relative to current flight path targets.
- Gathering and assessing relevant information from the world that increases understanding of the situation.
- Identifying appropriate, needed actions (whether monitoring actions or airplane control actions) for managing the flight path.

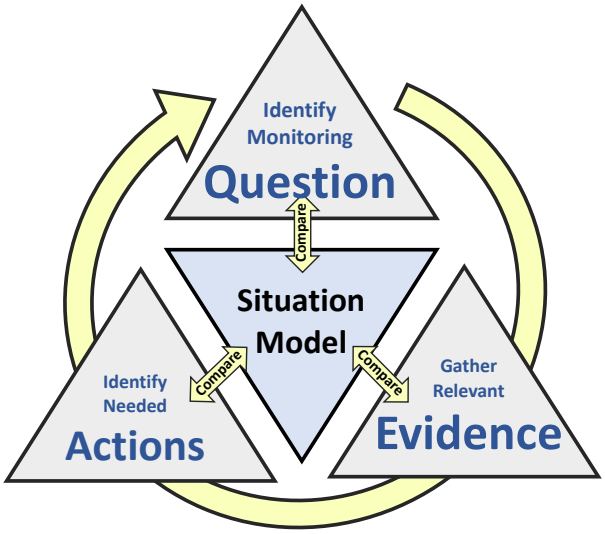

Figure 1. Situation model and sensemaking cycle.

These three activities all rely on the pilot's development of a "situation model," which is a mental representation that captures the current and upcoming state of the airplane and the operational environment. Indeed, it is the situation model that defines the success of sensemaking. We believe that improvements in monitoring can be gained by focusing on the development of this sensemaking skill and the skills that support and enable sense-making: task and attention management and flight crew communication. A question is how to provide that training and whether improvement in monitoring can take place following a relatively short training intervention.

For our study, we developed a short training intervention to investigate whether training on the sensemaking approach could aid a pilot in managing flight path-related monitoring challenges. Specifically, we hypothesized that a one-hour training session focused on applying a sense-making approach could improve pilot performance in identifying and resolving FPM-related challenges.

Another element of our study was to understand what situations are challenging for monitoring and, in turn, how we could use those challenging situations to assess monitoring performance. FPM can become challenging due to many types of events—e.g., transitions to inappropriate modes, shifts in the airplane's position relative to the desired flight path—and we were unsure how each type of event would affect monitoring performance. We needed an opportunity to construct a range of monitoring challenges to see which of them create more difficulties for pilots in a simulator setting. Measuring monitoring is challenging, in part because much of the process is cognitive, hence not

directly observable. Thus, it was important to identify exactly what would be scored to assess monitoring performance and changes in performance. We aimed to identify behavioral markers—and aircraft states—that reflect effective monitoring or its absence to enable effective evaluation.

A second experimental hypothesis was tied to the potential for the flight director (FD) to limit data gathering. Specifically, there is a concern during an autoflight approach that the pilot monitoring (PM) may focus too much on the FD. During an autoflight approach, the FD is locked on the glideslope and localizer signals. Some instructors prefer that the PM, who is not responsible for controlling to the flight path targets, scan a broad set of indications. If the FD is removed from the attitude indicator (ADI) on the PM's primary flight display (PFD), the PM may be more likely to scan beyond the ADI, e.g., out the window, at airspeed or altitude, at the flight mode annunciations, or the mode control panel, where modes are selected. Further, if there is a problem with the approach, such as a false glideslope, the PM should be able to detect that problem. Thus, we hypothesized that if the FD is not present (i.e., is off), the PM will monitor a wider or more diverse set of indications during an autoflight approach.

2. Method

2.1. Participants

The participants were 19 qualified first officers (FOs) who were active and current on the 737 NG (we attempted to recruit 24 pilots). Four of the participants were women. All participants were offered $100 and NASA stickers as an honorarium for participating. Participants were recruited, primarily through an email from the airline's pilots' union, to volunteer for a NASA study. Recruitment efforts targeted pilots who were in their first five years of employment at the airline; 17 of the 19 met the short-service target. We believed that these less-experienced pilots would be more likely to benefit from additional training. Median airline service time was 2.6 years; the mean was 3.3 years. This was a convenience sample and therefore cannot be assumed to be representative of the larger pilot population.

We gathered data on participant flight experience through an online survey and during the initial briefing phase of the study. Degree of precision and detail reported varied across participants. Table 1 shows participant flight hours. Median total flight hours was 7000 hours; median glass cockpit flight hours was 3000 hours (the 737NG is a glass cockpit).

Table 1. Participant Flight Time

	Mean	*Median*	*Range*
Total flight hours	8100	7000	4100 to 14000
Glass cockpit hours	3600	3000	400 to 8300
Pilot in command hours (n=12)	1780	1000	160 to 5000

Table 2 summarizes other aspects of participant experience. Participants typically had previous experience flying airplanes at a regional airline or in the military; two participants had both. Twelve of 19 had experience as pilot in command (PIC). We also created a distinct, more-subjective category of other "high-risk" flight experience either due to vehicle risk or type of operations. We asked participants to list the aircraft for which they were rated in addition to the one they currently

flew for the airline; the median response was three additional aircraft types. Participants often had instructor experience but this was not systematically reported.

Table 2. Prior Piloting Experience

Type of experience	*Number of participants (of 19)*
Regional airline	15
Military	6
Cargo	7
Pilot in command	12 (9 regional; 3 military)
Other "high risk" piloting	5

We hypothesized that diversity of experience might be associated with greater focus on monitoring and assessing the unfolding situation. We calculated a diversity score for each participant determined by the number of these experience categories: military piloting, regional piloting, cargo piloting, PIC role, and experience in our "high risk" type. Diversity scores ranged from 1 to 4 with a mean of 2.4.

Concerning education, all participants had completed (18) or were working on (1) a bachelor's degree; seven held a master's degree. Five pilots had a degree (bachelor's or master's) in a field not directly related to aviation.

2.2. Facilities and Equipment

The study was conducted at the training center of a major U.S. airline using their training rooms, simulator staff, and simulator.

2.2.1. Flight Simulator

The flight simulator was a CAE 737-700 full-flight simulator used in the airline's standard configuration (although the motion base was not used for the study). The simulator had a collimated out-the-window visual system. However, pilots were mostly flying in the clouds without any visual references. The simulator was set to record 231 variables (listed in Appendix A) with a sampling frequency of 6 Hz. The variables included flight variables, control inputs, and cockpit settings and alerts.

2.2.2. Eye-Tracking System

The simulator was fitted with a Seeing Machines single-crew configuration Crew Training System for the right seat. Crew Training System is a stand-alone single and/or multi-crew eye-tracking hardware and software package designed to support flightcrew training applications in the Full Flight Simulator (FFS) or Flight Training Device (FTD) environment. The system records crew visual scanning behavior across flight instruments and the cockpit environment. Specifically, 15 eye-tracker variables—including eye gaze, and eye-tracker diagnostics and timing—were collected at 60 Hz. The variables are listed in Appendix B.

The system uses a single camera and two separate infrared emitters/illuminators. The camera was placed above the FO's PFD with the two illuminators placed left and right of the camera. The

approximate locations of the camera and illuminators are shown in Figure 2. The system was calibrated prior to the study but not for each pilot individually.

Figure 2 shows the areas of interest (AoIs) that were defined in the eye-tracking software. Eight AoIs were defined: PFD; navigation display (ND); out-the-window (OTW) visual; mode control panel (MCP); electronic flight instrument system (EFIS) settings panel; flight management system (FMS); and upper and lower engine-indicating and crew-alerting system (EICAS) displays.

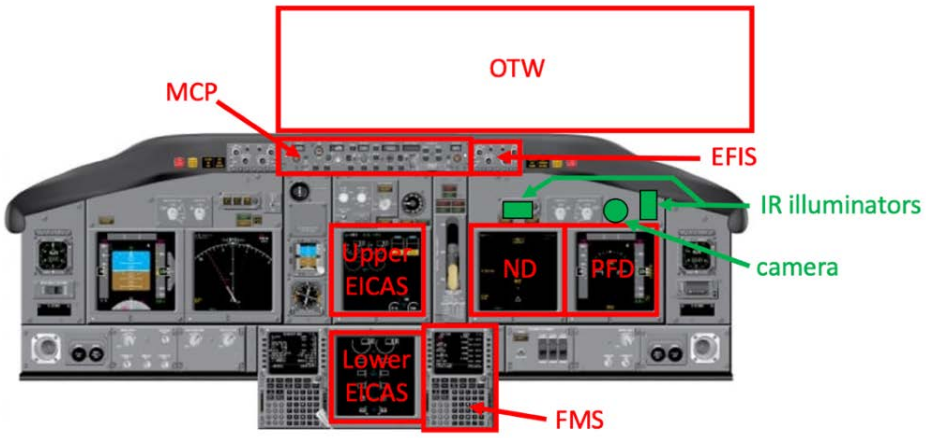

Figure 2. Areas of Interest defined for eye-tracking and location of eye-tracking equipment.

2.2.3. Video and Audio Capture

A single video camera was positioned over the left shoulder of the FO (who was also PM) and afforded a view of most of the flight deck interface, with more focus on the FO's instruments. It captured pilot conversation and activities but did not allow reading display values.

2.2.4. Timing

The data from the simulator, audio, video, and eye-tracking were all time-synched through a network time protocol (NTP) server. Both the simulator and eye tracker logged the synchronized universal coordinated time (UTC). The video was overlaid with the synchronized UTC.

2.3. Design

We used a 2 Training (pre- vs post-training, within subject) x 2 Scenario Order (Scenarios 1 & 2 first vs Scenarios 3 & 4 first, between subject) x 2 Display Configuration (with vs without FD, within subject) design as illustrated in Table 3. The Pre- versus Post-Training variable assessed the impact of training while Scenario Order was a counterbalancing factor. In Scenario Order 1, participants flew Scenarios 1 and 2 before training; in Scenario Order 2, they flew Scenarios 3 and 4 before training. Challenging events were nested in each scenario for a total of 15 and these 15 events were the items we scored. Dependent variables were derived measures from scoring pilot performance and pilot eye fixations on each event. Performance of each individual and on each item was assessed both pre- and post-training. The study also asked whether presence versus absence of the FD affected pilot monitoring.

Pragmatic factors of scheduling and simulator availability constrained feasible designs.

Table 3. Design: Pre- Post-Training x Scenario Block-Order (between subjects) x FD On or Off
(Assignment of Participants and Scenarios to Assess Effect of Training and of FD Status)

	Pre-Training	*Post-Training*
Subject Group 1 (Scenarios 1 & 2 Pre)	Scenarios 1 & 2	Scenarios 3 & 4
Group 1A	Scenario 1 On; Scn 2 Off	Scenario 2 On; Scn 1 Off
Group 1B	Scenario 3 On; Scn 4 Off	Scenario 4 On; Scn 3 Off
Subject Group 2 (Scenarios 3 & 4 Pre)	Scenarios 3 & 4	Scenarios 1 & 2
Group 2A	Scenario 3 On; Scn 4 Off	Scenario 1 On; Scn 2 Off
Group 2B	Scenario 4 On; Scn 3 Off	Scenario 2 On; Scn 1 Off

Each participant flew four operational scenarios: two scenarios before and two scenarios after training. Scenarios 1 and 3 and Scenarios 2 and 4 were "matched," that is, 1 and 3 had a similar set of monitoring challenges and 2 and 4 had a similar set of monitoring challenges.

The PM's FD was on (present) for two of the four scenarios and off (absent) for two of the four scenarios. The pilot flying's (PF's) FD was present for all scenarios. Table 4 presents the four combinations of scenario orders and assignment of FD status that were used to counterbalance these factors. Because we were unable to recruit 24 participants within the time window we had for simulator use, the orders were not evenly applied. Five participants were assigned to Orders 1, 2, and 3; only four participants were assigned to Order 4.

Table 4. The Four Configurations used to Counterbalance Scenario and Flight Director Status Order

Order 1	*Order 2*	*Order 3*	*Order 4*
Scenario 1 FD on	Scenario 1 FD off	Scenario 3 FD on	Scenario 3 FD off
Scenario 2 FD off	Scenario 2 FD on	Scenario 4 FD off	Scenario 4 FD on
Training	Training	Training	Training
Scenario 3 FD on	Scenario 3 FD off	Scenario 1 FD on	Scenario 1 FD off
Scenario 4 FD off	Scenario 4 FD on	Scenario 2 FD off	Scenario 2 FD on

Note this design confounds training and exposure to simulator sessions; the second set of scenarios are always preceded by exposure to the first two scenarios as well as by training. The resources made available for this study did not allow designation of a control group with no training between the first and second simulator session.

2.4. Procedure

A session for a single participant was scheduled for 3.5 hours, which was broken into five phases. The following describes the roles of the large experimenter team and then the five phases.

2.4.1. Experimenter Roles

The experimenter team used six roles to support data collection for each participant.

The Captain, who was always PF, and the Instructor Pilot (IP) were confederates recruited from the airline; were familiar with the research goals; and helped develop and script the scenarios. (The participant was not told that the Captain was a member of the research team.) They performed the following roles:

- *Confederate Pilot* (CP), who flew in the left seat as the PF. Three different airline pilots were trained to take turns in this role. Although the CP was the PF, he took a passive role regarding the monitoring challenges, relying heavily on prompts from the participant to fully manage the flight path. For example, the PF initiated a descent but did not initiate actions to ensure they would meet the waypoint crossing restrictions. The CP also made scripted, intentional errors, such as selecting a wrong mode or requesting an inappropriate flap setting (see Appendix C for details). Although CPs were briefed on how to respond to a variety of participant behaviors, there was some variation across sessions in how directly the participant had to identify a problem and recommend actions before the CP complied.
- *Instructor Pilot* (IP), who sat behind the flightcrew at the simulator controls and initiated and managed the simulator scenario. The IP also issued all ATC clearances. There were three pilots who could take this role. Although a scenario script was provided, some variability occurred across participants because of time shortages or researcher preference in wording.

Data collection was managed through these roles:

- *Simulator Manager* (SM), who ensured the simulator was functioning appropriately and managed two data-collection systems: audio/video and eye-tracking. All SMs were airline simulator technicians.
- *Observer* (Obs), who stood behind the participant and took notes on participant performance as each scenario was performed. All Obs were NASA Ames Research Center staff.
- *Data Manager* (DM), who input event markers for the beginning and end of monitoring challenges and ensured that all data were being captured. All DMs were NASA Ames Research Center staff.

The sixth role supported briefings and training:

- *Trainer*, who conducted orientation, training, and debriefing sessions (Phases 1, 3, and 5). The Trainers were NASA Ames Research Center staff. There were only two Trainers through the study to minimize training disparities among participants.

2.4.2. Procedure Phases

Each participant was guided through the following five phases.

2.4.2.1. Phase 1: In-Brief and Demographics (15 minutes)

The participant, Trainer, CP, and another experimenter met briefly in the lobby. The participant and Trainer then went to a briefing room for an individual briefing. The participant was given consent forms to review and sign (and offered a copy) and was then briefed on the structure of study activities. The study was described as NASA research on how pilot activities contribute to safety—monitoring was not mentioned. It was emphasized that the study in no way was an evaluation of the individual and that their individual performance results would not be shared. Demographic and

experience data were collected, reviewing and extending any information the participant had provided through an on-line survey done prior to that day.

Participants were told they would be flying descents to a landing; the specific approaches and airports were identified. They were offered the relevant Jeppesen charts for review, however, most participants opted to use their own charts on their iPads. Time available for review was short, typically about five minutes. Thus, the pre-flight briefing was much shorter than for normal simulator training. The CP and Trainer returned to the lobby at the same time that the participant arrived with another experimenter. They were escorted to the 737-700 simulator together to create the impression that both pilots were study participants, that is, the CP was not initially introduced as part of the research team.

2.4.2.2. Phase 2: Sim Session 1 (1 hour)

In this session, the participant and CP flew their first two scenarios. Participants were assigned to one of the four orders shown in Table 4, in the scheduling sequence. All participants were assigned to the PM role and sat in their usual, right-hand seat. The session started with introductions to the data collection staff (SM, Obs, and DM) and a very brief orientation by the IP. Each scenario started in cruise prior to the top-of-descent (T/D) point. To initiate a scenario, the IP told the flightcrew what arrival was being flown, which airport, and then ensured they had the appropriate charts. The flightcrew followed a script to ensure that the FMS had been correctly loaded. The IP asked the crew to conduct a descent briefing that included a risk assessment, which was standard practice for this airline. The CP led the short T/D briefing. Then the IP put the simulator in Run. The scenario was flown to either the initiation of a go-around or a successful landing at the airport, depending on the scenario.

After the initial scenario was flown, the IP re-positioned the simulator for the second scenario and the same sequence was used to set up and fly that scenario. When the second scenario was completed, the participant and CP left the simulator together but with different experimenters. The participant was then taken back to the original briefing room.

During each scenario, the Obs, in coordination with the IP, recorded when monitoring challenges were initiated and when they were resolved (if they were resolved). The Obs also took notes on how well the participant managed each challenge and about which specific actions were used to monitor and manage the flight path. Notes included key words or phrases expressed by the PM or an action taken that might have signaled being aware of a challenge, such as a hesitant gesture or a fixated look at a flight deck display without any comment.

Throughout the simulator sessions, we attempted to reduce variability from the two confederate roles to make the experimental condition as standard as possible across participants. Scenario events and CP actions were scripted and rehearsed. Indeed, two of the CPs were heavily involved in scenario development and familiar with the research goals. In particular, all CPs practiced being less active and less informative than would be the norm in a usual simulator training session. If the CP or IP began to explain too much, the Obs or DM cut them off. Some variation was required at times to respond to the participant. For example, when a participant noticed and was trying to understand an equipment failure, the CP or IP might provide a short explanation at the end of that scenario.

2.4.2.3. Phase 3: Training (1 hour)

Participants were offered a short break before or after the training session. When the training began, each participant followed a set of slides on a laptop computer that structured the information and

activities presented and the interaction with the Trainer. The training focused on monitoring skills as an active sensemaking process that builds up the pilot's model of the unfolding situation. The training session was designed to produce active learning and adapt to the participant's learning while providing standardized coverage, activities, and feedback across participants. More broadly, the intent was to produce a structured but open learning environment.

There were two versions of the slides; the participant was given the version that matched the airport flown in the Sim Session 1 scenarios (either KIAD/Dulles or KLAS/Las Vegas). The slides included a set of formatted questions that would appear in order to gauge the participant's understanding of the information presented in the slide. Participants voiced an answer when a question was encountered in the slides. Questions were designed both to check comprehension of the material just presented and to ask the pilot to elaborate or relate a concept to their prior experience, e.g., thinking of an example. The Trainer took high-level notes and recorded the participant's responses to activities and questions.

The structure of training activities remained the same throughout the study. However, slides were refined over the course of the study in several ways:

- Slide information density was reduced by editing out words and occasionally splitting content across two slides.
- Additional 'section header' slides were inserted to make the organization more visible.
- When alternative or familiar terms were introduced by participants in discussion, some were incorporated. For example, in discussion of the importance of communicating, one early participant exclaimed that they had learned "Don't be a secret keeper" and this phrase was added.
- A few comprehension check questions were dropped to reduce time.

Because participants were broadly assigned alternately to experimental conditions, impact of any improvements to the training procedure would impact conditions (specifically, scenario order) quite evenly.

During training orientation, the purpose of the study was described. Participants were told that in addition to investigating pilot contributions to safety we were interested in better training for monitoring for FPM and that next they would work through a "first draft" version of such training. The training progressed through the following six activities.

1. Self-Debriefing. (Note: The first three participants did not undergo this activity.) The participant was asked to type into the current slide a high-level debrief of the two scenarios they had just flown, describing any threat or unexpected events encountered and any mitigations taken. Additional prompts from the experimenter—suggesting they note whatever they remembered or thought interesting or important—were made in a few cases where participants were uncertain about what to do. Also, the Trainer typed for one pilot who was not comfortable typing.

2. What is Monitoring? The participant reviewed slides stating and diagraming the concept of monitoring for flight path management as "flying ahead of the plane" by gathering contextual understanding of the environment instead of just where you point your eyes.

3. Model of How to Monitor. The participant reviewed slides introducing a model of how to monitor. The slides introduced the concept of a Situation Model and the process of updating the Situation Model, which is a cycle of: 1) identifying an important question (the monitoring goal); 2)

getting and assessing relevant evidence; and 3) identifying any needed actions. This concept was shown graphically (Figure 1) and supported with text. This section had multiple comprehension and application questions. For example, they were asked to identify one important monitoring goal (What do you need to know?) from the scenarios they had just flown or to identify an important variable to check and the interface location where its current value could be found. At the end of this section, the participant was asked to draw the Situation Model diagram. To ensure and reinforce the participant's basic understanding of the steps in this cycle, we discussed their sketch, and provided feedback.

4. Communication. The participant reviewed slides with text and illustrations that emphasized the importance of communicating information from their Situation Model to the other pilot. These described when and what to communicate, the importance of keeping a shared Situation Model updated even when there is no problem, and the high value the airline places on communication. The participant recalled and reported examples from the scenario just flown regarding when they communicated about a problem and when they communicated even when things were fine.

5. Applying the Monitoring Model. This activity began with a review of the cycle of activity to update the Situation Model. The participant watched a short clip of another pilot flying a segment of the route just flown in the simulator with the goal of focusing on that PM's monitoring activities. After watching, the participant wrote a short paragraph to describe what the PM was doing and what they should have been doing. The participant was then asked to describe how to do it better, following the steps in the model update cycle and the resulting change in the Situation Model. Participants typed into a form to provide: 1) an example of a monitoring goal; 2) where to get the needed information and how to compare values; and 3) what actions would be needed. They were asked to summarize how this would update their Situation Model. They then identified particular types of information that would be important to communicate about this situation.

6. Prioritizing Monitoring Goal. Final slides provided guidance on general priorities in monitoring for FPM and how to monitor to produce an updated, shared understanding of the situation.

After training was completed, the participant and the CP returned to the simulator for the final two scenarios.

2.4.2.4. Phase 4: Sim Session 2 (1 hour)

The remaining two scenarios were run in the second simulator session using the procedures that were used for the first two scenarios (see Section 2.4.2.2). Upon completion of the final two scenarios, the participant was told that the CP was actually part of the research team and had been following a script that required him to be less helpful in FPM.

2.4.2.5. Phase 5: Final Debriefing (15 minutes)

The participant returned to the briefing room for a final debrief. Participants completed a questionnaire about the study. The questionnaire began with several rating scales asking about the value of and interest in the training, both overall and for different elements. This was followed by free-response questions about strengths and weakness of the training and simulator sessions and ideas for improvement. The Trainer reviewed the free-response questions to ensure the responses were comprehensible and to seek any relevant elaborations. As time permitted, the Trainer also answered any participant questions about the overall research effort and the study's purpose. We emphasized the purpose and importance of not talking about the study with other potential participants and obtained a verbal commitment not to discuss the study at all until 10 days after the

study completed. The Trainer then provided the participant with the paperwork necessary to receive a $100 honorarium and thanked them for their participation in the study.

2.5. Simulator Session Materials

The descent Scenarios 1 & 2 (Airport A) versus 3 & 4 (Airport B) were counterbalanced to be presented as the Pre- or the Post-Training session. To measure monitoring performance, 15 challenging events were designed so that noticing and understanding the event would lead to specific, identifiable behaviors, and enable objective scoring (Table 5). Behaviors were typically talking to the PF but some were control actions. The response might be communicating with the PF, taking an action (such as lowering flaps), or both (such as communicating with PF and then calling ATC to request relief). Thus, each monitoring challenge had associated appropriate behavior to use as a standard for scoring the observed behavior. The set of behaviors for an event can be considered a graded behavior marker. Appendix C describes all four scenarios and their scoring rules in detail.

Integrating these challenges for the PM to catch while maintaining realism relied critically on collaboration with senior pilots, drawing on reported safety events and their own line experience. The challenging events in Scenarios 1 & 2 versus 3 & 4 were designed in pairs to pose challenges of similar types and difficulty but in different airports and conditions. Matched pairs proved possible for 14 of the 15 events (Table 5).

Table 5. Challenge Event Descriptions, by Matched Pair (where possible)

Challenge Type	*Scenario 1*	*Scenario 3*
High on path (ATC)	#1 Slowed by ATC	#9 Held height by ATC
Inappropriate mode	#2 PF remains in VNAV	#10 PF selects HDG SEL
Instrument issues	#3 Given wrong altimeter setting	#11 False glideslope
Did not enter value	#4 Field elevation not set on MCP	
	Scenario 2	*Scenario 4*
Inappropriate mode	#5 Auto-flight/PF interaction VS	#12 PF engages LVL CHG
Shortened lateral path	#6 ATC gives direct-to	#13 ATC gives direct-to
Inappropriate mode	#7 PF selects LNAV	#15 PF fails to arm APP
Airspeed error	#8 PF calls flaps 25 when too fast	#14 PF fails to call for flaps 5

2.6. Dependent Measures

For each scenario, we collected video and audio of each participant's performance, took notes on the participant's performance, and captured their eye fixations. We derived three measures for each challenging event. Two closely related measures, the Performance Score and Success/Fail Score, were based on reviewing audio/video and airplane performance to assess how well the participant identified and resolved each monitoring challenge event. The Performance Score provided an ordinal rating of how well the PM responded.

- *Performance Score.* The intended scoring was to code four operationally distinct performance levels: two passing, two unsuccessful. A 3-level scale was used for five and a 2-level scale for two events because these events had fewer operationally consequential levels of performance. These scores were assigned for every event that was presented as intended (we excluded a few missing trials or trials where the

participant or experimenters pushed events off their intended course). Unscorable events produced missing data. Conceptually, the levels were as follows:

  - Bad (1): Clearly unacceptable performance, often with operational consequences, e.g., overspeed the flaps, misses approach, misses altitude restriction.
  - Undesirable (2): Performance is less than would be expected in an evaluation setting but there are no significant operational consequences, e.g., fails to use the optimal mode but still makes altitude restrictions or fails to coordinate with ATC.
  - Less than ideal (3): Typically falls short of ideal performance in some way but captures the most important elements of performance and indicates awareness of the challenge, e.g., performance occurs later than it should but there are no operational consequences.
  - Ideal (4): Performance as prescribed by the scenario developers, e.g., ideal autoflight mode; good coordination with ATC; meets flight plan restrictions.

  Details of scoring criteria for each challenging event are in Appendix C.

- *Success/fail.* This score was a binary scale, collapsing the four-point Performance Score. Scores of 1 and 2 were "fails" while 3 and 4 were "successes."
- *Time to Successful Resolution.* The third measure was assigned only for successfully managed events. This measured the time from the onset of the challenge (a critical change that required monitoring and should elicit an action) to its successful resolution.

Several dependent measures were derived from the eye-tracking data. Fixations and saccades were identified from the eye-tracker data using a simple velocity-based algorithm. Using the distance of the pilot to the AoI, the point-to-point velocity between two gaze points was calculated. Each gaze point was then classified as a fixation if the velocity between points was below 100 deg/s and a saccade if it was above this threshold (Salvucci & Goldberg, 2000).

Next, the following measures were calculated:

- *Dwell Time* is the time a participant's gaze remains on a given AoI. It was calculated by aggregating time while the gaze vector intersected with a certain AoI during a monitoring challenge for each of the following AoIs (see Figure 2):
  - Primary flight display (PFD)
  - Navigation display (ND)
  - EFIS control panel (EFIS)
  - Mode control panel (MCP)
  - Out-the-window view (OTW)
  - The display for the flight management system (FMS)
  - Upper EICAS
  - Lower EICAS
- *Proportion Dwell Time* for a specify AoI was its Dwell Time as proportion of the total duration of the challenge for a specific pilot. As opposed to Dwell Time, Proportion Dwell Time controls for variation in how long the relevant flight segment took.

- *AoI Neglect Latency* is the time between moving fixation away from an area of interest and again fixating that same area as determined by the intersection of the gaze vector with an AoI. This provides a measure of how recently information at that AoI might have been sampled.

We also explored two additional measures: average fixation dwell time and dwell time calculated for elements within the PFD.

2.7. Data Processing

A web-based tool written in Angular was specifically designed for this study to support data integration and visualization of all three data streams. The tool is described in Appendix D. The tool provided animated playback of the simulator data displayed on a representation of the flight deck instruments; it synchronized display of this data with eye fixation sequences projected onto the eye-tracking areas of interest (Figure 2) and with the over-the-shoulder video of pilots' activities. The first step in aligning the data was to resample the simulator data to the sampling rate of the eye-tracking data from 6 Hz to 60 Hz using the synchronized UTC time in both data streams. Next, the video data were aligned in the software tool by applying an offset based on the difference between the UTC time overlaid on the video and the UTC time in the simulator/eye-tracking data.

The start and stop times for each challenge were determined by playing back the data for each scenario using the software tool and by analyzing the simulator data in MATLAB. Start times for each challenge were defined by the time the PF or the IP introduced the error (e.g., giving an erroneous instruction to the PM) or at a specific point in the flight plan (e.g., T/D). The challenges ended when the PM verbally expressed concerns or performed an action to solve the error or at a specific point in the flight plan.

The eye-tracking data from two pilots were not used in the analysis. The eye tracker software produced a quality measure that was significantly lower for these pilots. In addition, the eye tracker was not able to detect where the pilot was looking for large portions of the scenarios. In addition, one pilot did not complete Scenario 4 of the experiment due to time constraints.

3. Results

3.1. Performance Data

Table 6 shows the scoring by event: 285 data points (19 participants *15 events) were planned and we had 18 missing data points. Data were missing either because the participant did not get an opportunity to manage the event or because the participant had an unexpected scheduling conflict. Missed opportunities resulted from the scenario being adjusted in a way that either removed or shortened the opportunity to present the monitoring challenge. Although each event was designed to have a 4-level score, seven events were only sensibly scored on a reduced scoring level, two events on only two levels, and five events reduced to three levels. Dropped scoring levels are marked NA in Table 6.

Table 6. Performance Scores and Missing Data Tallies for Each Event

Event	*Event ID*	*Missing Data*	*Bad (Score = 1)*	*Undesirable (Score = 2)*	*< Ideal (Score = 3)*	*Ideal (Score = 4)*
Scenario 1 Event 1	1	1	4	3	1	10
Event 2	2	3	3	8	1	4
Event 3	3		5	5	6	3
Event 4	4		6	1		12
Scenario 2 Event 1	5	1	3	1	xxxxx	14
Event 2	6	2	4	xxxxx	xxxxx	13
Event 3	7	1	6	xxxxx	xxxxx	12
Event 4	8		8		1	10
Scenario 3 Event 1	9	6		2	0	11
Event 2	10			3	xxxxx	16
Event 3	11		1	6	3	9
Scenario 4 Event 1	12	1	9	xxxxx		9
Event 2	13	1	3		2	13
Event 3	14	1	6	xxxxx	4	8
Event 4	15	1		xxxxx	14	4
Totals (of 285)		**18**	**58**	**29**	**32**	**149**
% of 285	***285***	***6.32%***	***20.35%***	***10.18%***	***11.23%***	***52.28%***

Performance on all three measures was highly variable across both events and participants. Table 7 presents two perspectives on the data, showing performance both with each participant weighted equally and with each event weighted equally. That is, the participant scoring treated the participant as the unit of analysis and the event scoring treated the event as the unit of analysis. For example, these perspectives show that there was an item that was judged correctly by 100% of participants, but there was no participant who was correct on all items. There are different numbers of missing data in different cells, so mean scores will differ depending on how the data are grouped. Maximum and minimum scores will be even more influenced. Thus, it is reassuring to see a similar pattern from each perspective. Overall, participants were successful on 68% of the events with a 64% range (29%–93%). Performance on the four-point performance score was also highly variable with an average score of 3.0 (out of 4). Variability across events was also high, with 100% success on two events, and with three events having 50% or less success.

Recall that Scenarios 1 and 3 and Scenarios 2 and 4 were "matched"—1 and 3 had a similar set of events and 2 and 4 had a similar set of events. We looked at the order of presentation: Scenarios 1 & 2 presented first (before training) or Scenarios 3 & 4 presented first (before training). Table 7 shows performance broken down by Scenario order. Challenge events in Scenarios 1 & 2 tended to be more difficult than those in Scenarios 3 & 4, suggesting that the sets with analogous structure were not as closely balanced in difficulty as ideal. In addition, participants seeing Scenarios 1 & 2 first tended to do less well.

Table 7 also shows completion/resolution time (in seconds) for events that were successfully completed (score 3 or 4); generally, short times are good. While more difficult challenge events might be expected to take longer, these more difficult items might be more likely to fail and thus not be included in this measure. Similarly, better-scoring individuals were more likely to succeed on difficult as well as easy items. In short, different samples of events are included in different groupings of data and thus, care is needed in interpreting this measure.

Table 7. Two Perspectives on Event Performance: By Pilot and by Event

	% Success	*Score (SD)*	*Time to Success (seconds)*
Each Pilot Equal Weight			
Overall			
Mean	**68%**	**3.01(.60)**	**97**
Max	93%	3.80	179
Min	29%	1.79	28
Pilots with Scenarios 1 & 2 First			
Mean	61%	2.81	84
Max	73%	3.53	179
Min	29%	1.79	28
Pilots with Scenarios 3 & 4 First			
Mean	76%	3.21	112
Max	93%	3.80	132
Min	29%	2.07	88
Each Challenge Event Equal Weight			
Overall			
Mean	**69%**	**3.02(.43)**	**105**
Max	100%	3.69	393
Min	31%	2.37	17
Events in Scenarios 1 & 2			
Mean	60%	2.88	107
Max	78%	2.37	393
Min	31%	3.39	17
Events in Scenarios 3 & 4			
Mean	78%	3.20	102
Max	100%	3.69	171
Min	50%	2.50	30

In summary, our event set overall did not suffer from floor or ceiling effects. Both pilots and events are important sources of variance. Importantly, each event and each pilot were measured before and after training. Thus, the effect of both item and participant can be addressed in an integrated, mixed model assessing the effect of training.

3.2. Effects on Performance Score Pre- vs Post-Training

Table 8 provides an overview of Performance Score Pre-Training and Post-Training. At the individual level, 13 of the 19 participants improved; one kept the identical score; and 5 decreased.

Table 8: Performance Score Before and After Training

Subject Group	*Pre-Training*	*Post-Training*
G1: scenarios 1 & 2 first	2.51 Items 1–8	3.19 Items 9–15
G2: scenarios 3 & 4 first	3.17 Items 9–15	3.28 Items 1–8
Overall pre- and post-training average	2.80	3.23

Figure 3 illustrates difference in performance on each event before and after training with 13 of the 15 events improving. To assess the impact of training on the ordinal, four-level Performance Score variable, we used the mixed-effects model for ordinal data in the clmm function (ordinal package in R version, Christiansen 2019). Model assessment is carried out in three steps. A model including the factor of interest is assessed. An otherwise identical model without the factor of interest is assessed. An ANOVA procedure assesses the probability that the fit provided by the larger model, which includes the factor of interest, was an improvement over the fit provided by the simpler model, taking into account the larger degrees of freedom in the larger model. If the probability is low that the larger model improved the fit by chance, the impact of additional factor is significant.

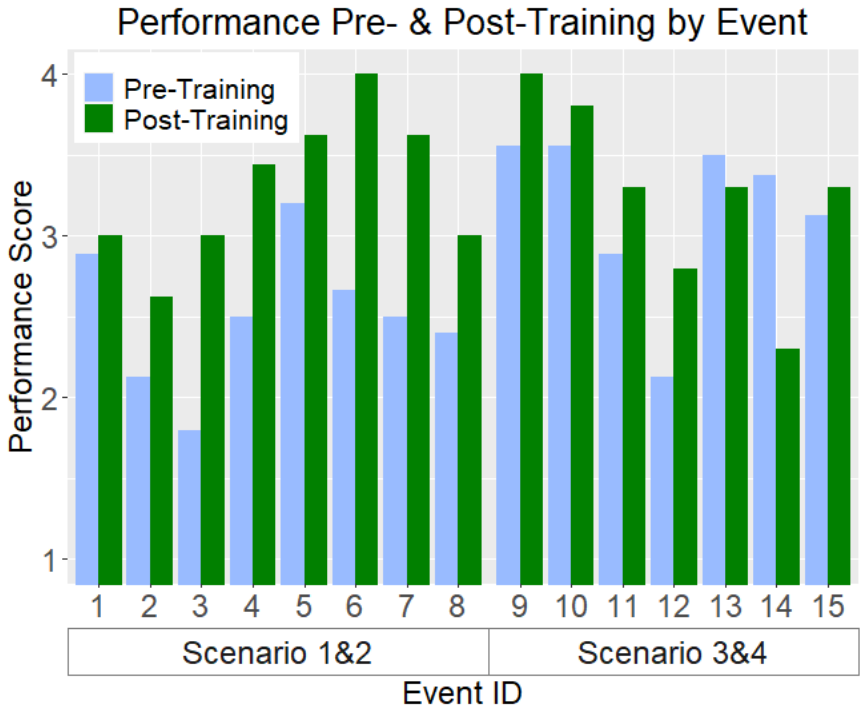

Figure 3. Performance on each event Pre- and Post-Training.

We modeled Performance Score (ordinal) as predicted by Training (fixed factor), by Scenario Order (fixed factor), by the Training X Scenario Order interaction, by Participant (categorical random factor), and by Event (categorical random factor). We compared this to a model that did not include the Training factor.

An ANOVA comparing the two models found that the model with the additional factor of Training provided a significantly better fit, $\chi2(2) = 10.868$, $p = .00437$ (likelihood ratio tests of cumulative link models). Analogous comparisons for effect of Scenario Order or the Order X Training interaction did not find a significant contribution for either (for Order, $\chi2(2) = 2.066$, $p = .356$, or for Order X Training, $\chi2(1) = .624$, $p = .430$). Pre- and post-training scores differ significantly. Note that while an interaction is suggested by the data, this was not significant. We consider additional clues after looking at additional performance variables.

An additional perspective on performance comes from looking at the correlations, across participants, on several participant-level, average scores. We derived each participant's average Performance Score, average score on Pre-Training events, average score on Post-Training events, and the Change Score between Pre-Training and Post-Training scores. The three Spearman correlations among overall, Pre-Training, and Post-Training Performance Scores are all positive and significant, as would be expected. Turning to the Change Score, the Pre-Training Score is correlated negatively and significantly with the Change score $r = -.570$, $p = .011$: people who did poorly initially showed more improvement.

3.3. Effects on Successes of Pre- vs Post-Training

We tested for the effect of training on the proportion of successes, using the same logic as described for Performance Score. Since Success is a binary variable, we used the generalized linear model (glmer function) and specified the distribution family as binomial (package = lme4 v1.1-26; Bates et al., 2015). As before, we made ANOVA comparisons between two models to test whether addition of a model component significantly increased fit. This is worth testing separately from overall score as it could be possible for differences to result from less important improvements within successes or within failures (scores of 4 rather than 3 or 2 rather than 1) rather than from more consequential increases in number of successes. While a less sensitive measure, number of successes pulls out a particular, important type of change.

Inclusion of Training in the model provided a significantly better fit, ($\chi2(2) = 10.341$, $p = .00568$. Analogous comparisons for effect of Scenario Order or the Scenario Order-Training interaction did not find a significant contribution for either Scenario Order, $\chi2(2) = 4.999$, $p = .0821$, or for Order X Training, $\chi2(1) = 3.066$, $p = .07999$. While similar to the findings with the Performance Score, the effects of order and the order X training interaction might be considered marginal for success score, given their $p = .08$.

Tests described above are quite sensitive because they simultaneously measure the variability from subject and from event effects. Another perspective can be used to assess the robustness of the Training effect: looking at the difference by item and the difference by individual. The 15 events can be ranked by how much each event's score changed pre- and post-training and tested for whether the differences are greater than chance. Using the Wilcoxon Signed Rank test, the statistic $W = 114$ shows a probability much less than $p < .01$ that the difference is due to chance. A related (nonparametric) test, just judging each item as a "success" if there is a positive change and a "fail" otherwise also shows a significant difference, $p = .004$ on the binomial test.

Turning to the difference by individual, the 19 individuals can also be ranked by how much each person changed (the difference score) and these differences compared to a chance model if there were no pre-/post- difference. Testing again with the Wilcoxon Signed Rank test finds W = 171 and the probability of a difference score this large is p <.01. Performance was scored as a "success" if there was improvement and a "fail" otherwise (including 1 individual who did not change in the fail category), with 13/19 successes. Here the binomial test falls short of the significance level of p <.05, with only a marginally significant value, p = .08, in contrast with the Wilcoxon comparison.

3.4. Effects on Time-to-Success Pre- vs Post-Training

Table 9 shows time to complete an event successfully. If an event was not completed successfully, it did not have a Time-to-Success score. Times-to-Success values were similar Pre- and Post-Training, with successful events in Scenarios 1 & 2 tending to be completed faster than in Scenarios 3 & 4. The Time-to-Success variable is fit by the Gamma distribution and using the glmer function with family = Gamma showed no improvement of model fit when Training was included in the model. Figure 4 shows the time distributions pre- and post-training by event, and illustrates the item differences in variability of Time-to-Success. Figure 5 shows distribution of Time-to-Success, with Pre-training shown in the top panel and Post-test shown in the bottom.

Table 9. Completion Time for Events with Successful Outcomes

Subject Group	*Pre-Training*	*Post-Training*	*Totals*
Scenarios 1 & 2 first	81.9 (SD = 99) n = 35	85.7 (SD = 76.0) n = 51	84.2 (SD = 85.6) n = 86
Scenarios 3 & 4 first	120 (SD = 98) n = 45	108.0 (SD = 138.6) n = 52	138.6 (SD = 120.9) n = 97
Totals	103.3 (SD = 99.5) n = 80	97.0 (SD = 112.1) n = 103	99.7 (SD = 106.5) n = 183

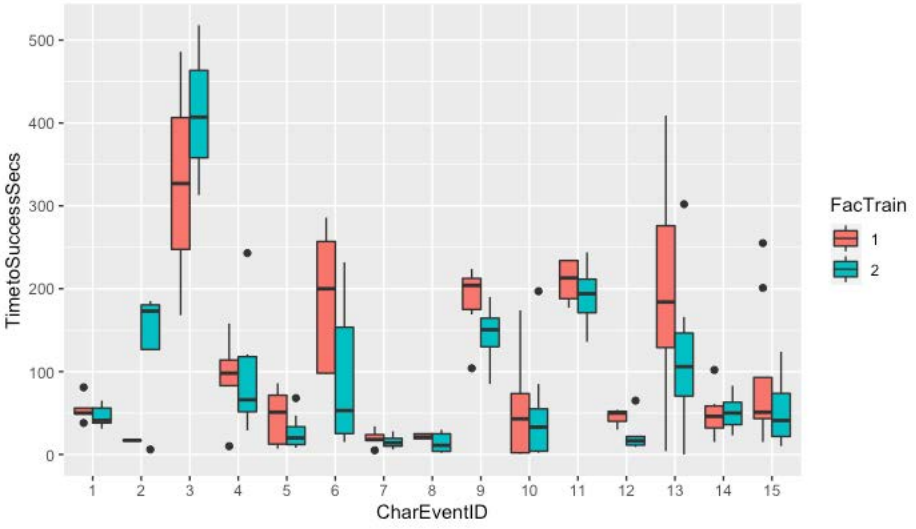

Figure 4. Time-to-Success Pre-Training (red) and Post-Training (green) in seconds.

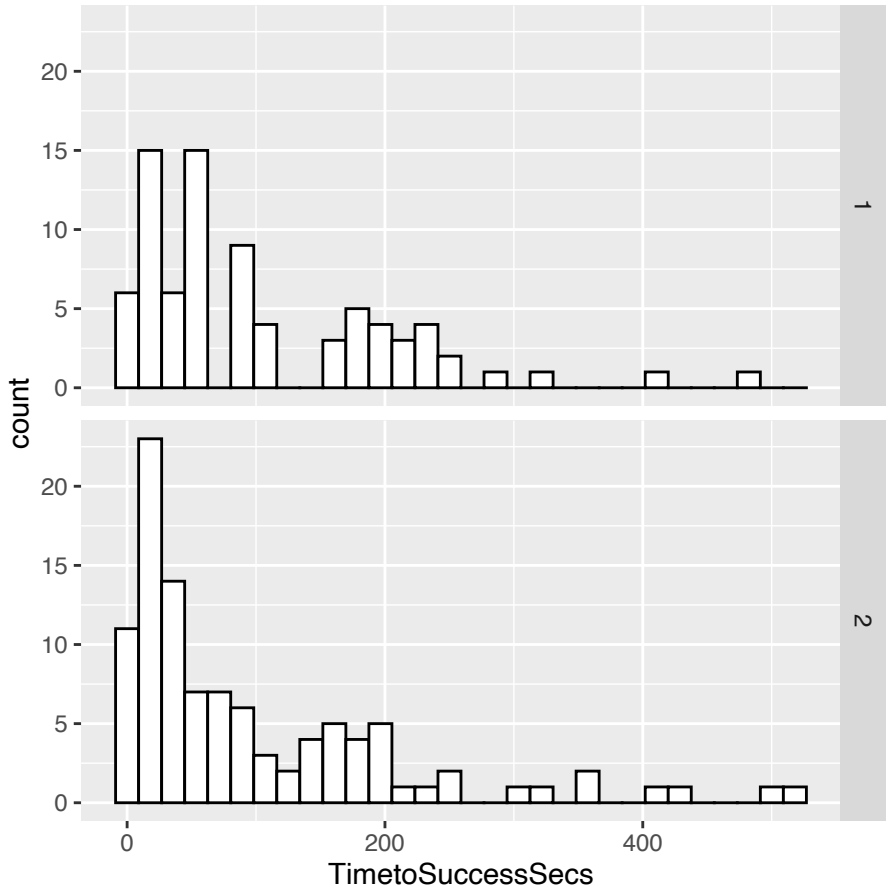

Figure 5. Distribution of Time-to-Success Pre-Training (top panel) and Post-Training (bottom panel) in seconds.

The Time-to-Success data could be explored in more detail, as the meaning and importance of timing differs across items. However, this event-focused analysis was not a priority and was not pursued.

3.5. Exploration of Performance: Demographics

Throughout the analyses conducted, the only significant effects come from the training factor, not order of presentation. The order factor was intended as a counter-balancing factor, which ideally would have no effect. However, the change from pre- to post-training performance tends to be greater for the events in Scenarios 1 & 2 and for the participants who saw these events first. Thus, we might explore whether we can identify any demographic variables that predict performance prior to training and if so, whether these tended to be distributed differently between the two groups of participants.

We collected quantitative, demographic data on seven variables described in the methods section and shown in Table 10. Table 10 shows Spearman correlations of these variables with the Performance Score on the Pre-Training block. None of these variables correlated significantly with performance on the Pre-Training block. The correlations for Flight Hours and Glass Hours were marginally significant, with $p < .09$. Surprisingly, the correlations were negative, such that more experience was associated with poorer performance on the initial, Pre-Training events. Number of glass cockpit hours showed the strongest, though nonsignificant, correlation with initial performance. In short, we do not have strong experience measures that predict performance and the

predictive direction is negative; the more experienced pilots showed nonsignificant tendencies to do less well on the Pre-Training block. Lack of a positive association (and presence of negative trends) surprised us, and it remains an open question whether and how experience and monitoring these types of events are related.

Table 10. Spearman Correlations of Demographic Variables with Performance Pre-Training

Flight Hours	*Glass Hours*	*Diversity*	*Military (Y/N)*	*121 PIC Hours*	*PIC Mil Hours*	*Years at Airline*
-0.41	-0.42	-0.07	-0.07	0.00	-0.39	-0.21

Table 11 shows values of these demographic variables for the two groups of participants. Despite the lack of easy-to-interpret or significant correlations between these variables and performance, we can look at similarities and differences between the two groups. Most of the measures are similar between the two groups. Participants who did Scenarios 1 & 2 pretraining tended to have more flight hours; whether this might lead to better (prior expectation) or worse (no significance correlation in this direction) performance is unclear.

Table 11. Demographic Variables of the Two Subject-Groups

Group	*Flight Hours*	*Glass Hours*	*Diversity*	*Military (Y/N)*	*121 PIC Hours*	*PIC Mil Hours*	*Years at Airline*
Scenarios 1 & 2 first	8491.1	4925	2.5	1.3	604	455.56	2.81
Scenarios 3 & 4 first	7664.44	2456.67	2.33	1.38	906.67	500	3.76

In short, our demographic measures are not strong predictors of performance. The measures do not provide a basis for predicting that participants seeing Scenarios 1 & 2 first might be expected to do worse.

For events, unlike participants, we have no data independent of performance to predict event difficulty. Initial difficulty does affect how much events can improve.

3.6. Effect of Flight Director

3.6.1. Effect of Flight Director on Performance

The presence or absence of the FD was hypothesized to affect gaze, particularly on Approach. No effect on performance was predicted and none of the challenge events took place during Approach. Nevertheless, we compared performance on challenge events with and without the FD. With the FD on, the mean overall performance score was 3.02 (SD = 1 .22) and with the FD off, the mean was 3.05 (SD = 1.24). We ran clmmr models with Training, FD status and their interaction as fixed factors, and Participant and Challenge Event (as categorical random factors) and compared this to a model that did not include the FD status. There was no effect of including FD in the model, and the model without it had a slightly better fit metric (lower AIC score). Thus, there was no evidence that FD status influenced performance, when assessed at this broad level.

3.6.2. Effect of Flight Director on Eye-Tracking Measures

A key hypothesis guiding the study was the possibility that, in the final approach, having the FD on would draw an inappropriately large share of attention (higher Proportion Dwell Time) to the PFD and, more specifically, to the FD, and reduce attention elsewhere, possibly impacting awareness of other elements of the approach. A prediction about the AoI Neglect Latency (ANL) might be seen as an extension of the prediction about Proportion Dwell Time in the original hypothesis. That is, if FD off meant pilots looked around more, this might suggest that it would take pilots longer to return their gaze to the PFD and PFD ANL would be shorter.

Proportion Dwell Time (PDT) on the PFD with the FD on or off is shown in Table 12. The table shows that the trend was the opposite of the predicted direction; that is, a higher PDT was observed when the FD was off, not when on. Using the described linear mixed model analysis approach, we tested whether PDT on the PFD differed significantly with FD on versus off. The trend observed in Table 12 was marginally significant ($\chi2(1) = 3.65$, $p = 0.056$). In sum, there was a non-significant trend in the opposite direction to that hypothesized, with a longer PDT on the PFD when the FD is off.

Table 12. PFD Proportion Dwell Time with the Flight Director On or Off in the Final Approach

	Proportion Dwell Time (sec.) Mean (Standard Deviation)		
	Scenario 2	*Scenario 4*	*Mean by FD Status*
FD off	0.51 (0.08)	0.47 (0.11)	0.49 (0.10)
FD on	0.44 (0.10)	0.38 (0.13)	0.41 (0.12)
Mean by scenario	0.48 (0.10)	0.43 (0.13)	0.45 (0.11)

Turning to the effect on ANL, Table 13 shows the average neglect latency with and without the FD. The neglect latencies were similar with longer gaps before returning to the PFD when the FD was on (not when off, as predicted). Note that because there were many AoIs, the time spent fixating one area (PDT) and the gap between fixations (ANL) are measuring very different things: one could be high and the other low, or vice versa. The effect of FD on the PFD AoI Neglect Latency was tested by comparing the fit of a model with the FD factor included versus a model without the FD factor. The procedure of the statistical test was the same as for PDT. The mixed model analysis found the effect of the FD on Neglect Latency was not significant ($\chi2(1) = 1.23$, $p = 0.27$).

Table 13. PFD Neglect Latency with the Fight Director On or Off in the Final Approach

	AoI Neglect Latency (sec.) Mean (Standard Deviation)		
	Scenario 2	*Scenario 4*	*Mean by FD Status*
FD off	1.82 (0.37)	2.31 (0.52)	2.08 (0.51)
FD on	2.14 (0.61)	2.42 (0.58)	2.27 (0.59)
Mean by scenario	1.98 (0.52)	2.36 (0.53)	2.17 (0.55)

The particular scenario of the four used (the Scenario variable) was not hypothesized in advance to affect either Dwell Time or ANL. Therefore, the inferential status of significance testing for either

PDT or ANL is not the same as testing for the effect of FD status. However, the relatively large contribution of Scenario to the models for each dependent measure prompted assessing the effect of Scenario on each. We used the same analysis procedure as for assessing the effect of FD status. The effect of Scenario was significant and very similar for both PDT ($\chi2(1) = 5.73$, $p = 0.018$) and Neglect Latency ($\chi2(1) = 5.74$, $p = 0.017$). Lower PDT and higher ANL were found for Scenario 4 (approach to KLAS) compared to Scenario 2 (approach to KIAD).

This suggests that there was systematic variability in eye-tracking measures in our data and thus that our data were sensitive enough to detect some effects. In turn, this suggests that the lack of significant effects of FD status was not simply due to very noisy data.

3.7. Eye Tracking Differences between Successful and Unsuccessful Performance

It is not unreasonable to ask whether eye tracking patterns might be different on successful versus unsuccessful performance, at least for events where success depended on knowing the value of a specific variable that is best read from a specific display. Four of the challenge events seemed to have this property. Two scenarios concern way point restrictions and would benefit from using the vertical situation display (VSD) within the NAV display, namely Scenario 1 Challenge Event 1 (S1C1 = Challenge #1 in Table 5) and Scenario 2 Challenge Event 4 (S2C4 = Challenge #8 in Table 5). Two scenarios required airspeed information, presented on the PFD, namely Scenario 4 Challenge Events 3 & 4 (S4C3 = #14 and S4C4 = #15 in Table 5). Note, however, that success might be associated with more time spent on the most relevant AoI (ensuring an accurate, current awareness of the variable) or with less time spent (because successful pilots are more efficient or strategic in their sampling).

Figure 6 provides PDT for all eight AoIs on each of the 4 challenges, for each of the 17 pilots with eye tracking data. The bars for a particular pilot do not necessarily add up to 100% as pilots may look at none of the AoIs for some portions of an event and eye-tracking data might not be available. Means per AoI across pilots are provided by red dashed lines. The plots on the left provide results for pilots who successfully completed the challenges, while the plots on the right provide results for pilots who were unsuccessful. This division explores whether successful pilots might allocate their gaze in a more structured or efficient way than unsuccessful pilots allowing them to detect problems more quickly. Note that more pilots were successful than unsuccessful and the number of unsuccessful pilots is particularly small for some challenges. This means one should be cautious in drawing conclusions from the differences between successful and unsuccessful pilots presented here. Even though pilots performed challenges before and after training, results were not subdivided accordingly as the training did not focus on changing pilots' scan patterns or gaze allocation and initial inspection did not reveal differences in eye tracking before and after training.

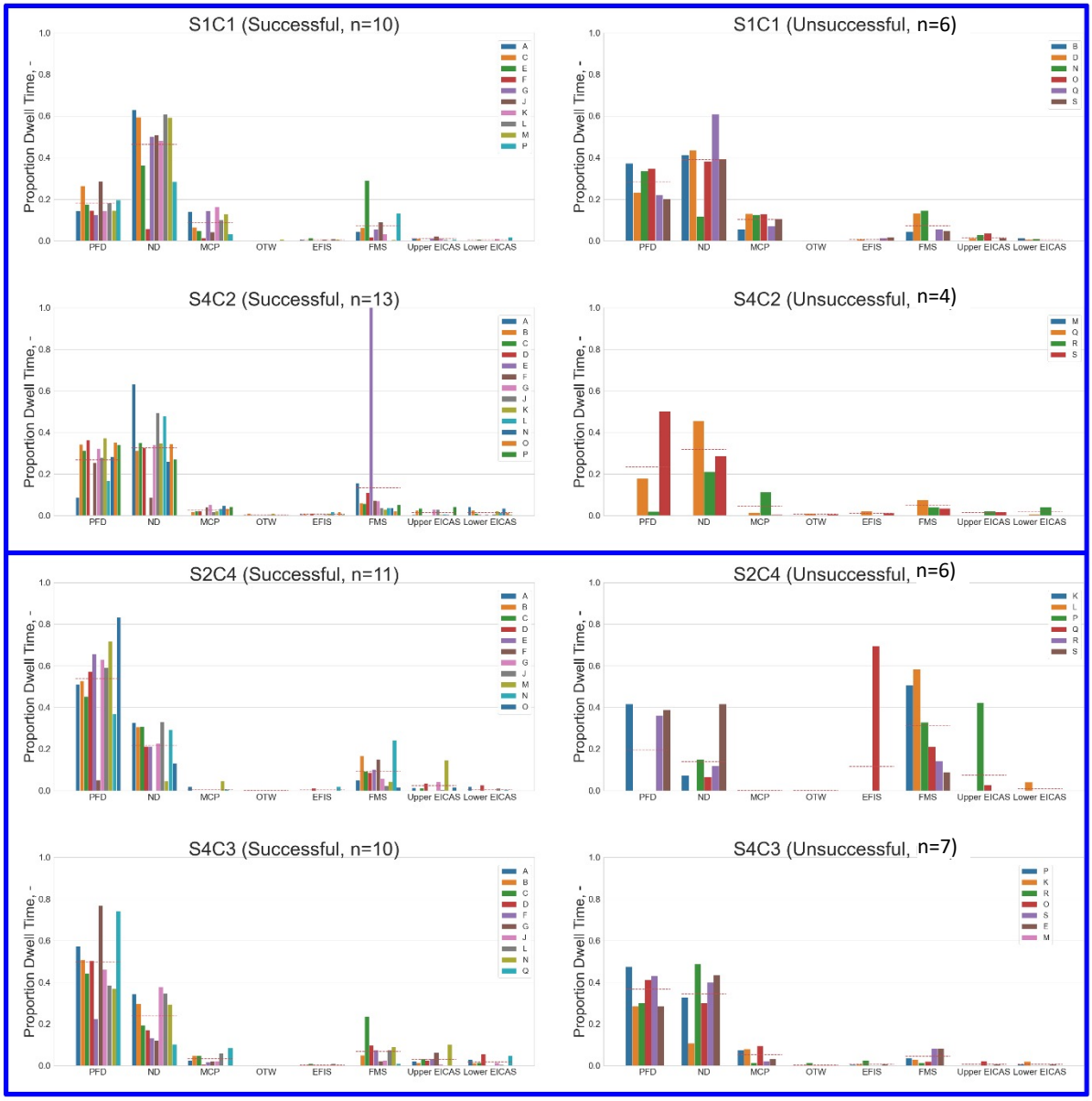

Figure 6. Dwell time by individual participant (indicated by color) grouped by successful or unsuccessful outcome. Data for the four monitoring challenges where AOIs most relevant to success might be identified. S1C1 and S4C2 are challenges to meet waypoint restrictions; here altitude is critical to track and so the vertical situation display on the ND might be particularly helpful. S2C4 and S4C3 are challenges to extend the flaps at the appropriate time; here vertical speed is critical to track and so the VS display on the PDF might be particularly helpful.

The top four plots in Figure 6 provide PDT for the two challenges with waypoint restrictions. Most of the pilots' attention was on the PFD and ND. Pilots had the use of a VSD on the ND providing information on their vertical flight path. This is a valuable source of information in the waypoint-restriction challenges and, as such, we were anticipating higher Proportion Dwell Times for the ND than the PFD. Figure 6 indicates mean PDT was, indeed, much higher on the ND compared to all other AoIs for Scenario 1, Challenge 1 (S1C1), and particularly for pilots who were successful.

However, this is not the case for Scenario 4, Challenge 2 (S4C2). S4C2 had a more demanding airspeed restriction (available on the PFD) which might cause a higher PDT on the PFD and a more even allocation between PFD and NAV, relative to that in S1C1. Thus, the scenarios specifically selected for the simplest predictions about PDT proved more complicated than anticipated. Additional detail about why this might be the case based on pilot-aircraft performance is presented in Appendix C.

PDTs were relatively low for all other AoIs. In one case for S4C2, the PDT for the FMS was 100%. This means that for the entire duration of the challenge the pilot was looking at the FMS. Note that PDT does not imply a length of time so it could have been the challenges for these pilots were shorter. Inspection of the data suggests successful pilots may spend more time on the AoIs most relevant to complete the challenge; that is the ND for S1C1 and the PFD and ND for S4C2.

The bottom four plots of Figure 6 provide PDT for the challenges concerning flap extensions. For these challenges, flap extension was restricted by airspeed and thus monitoring airspeed on the PFD was most important. Figure 6 indicates mean PDT was highest on the PFD compared to all other AoIs for both Scenario 2, Challenge 4 (S2C4) and Scenario 4, Challenge 3 (S4C3). The ND is the second-most-focused-on AoI for pilots who successfully completed the challenge, followed by the FMS. Inspection of successful and unsuccessful pilots suggests that unsuccessful pilots focused on the PFD less. Furthermore, unsuccessful pilots focused on the FMS much more in S2C4 and one unsuccessful pilot more on the EFIS settings.

Looking at Figure 6, an interesting observation can be made. Dwell time patterns seem to differ more across scenarios for successful trials, while patterns show less differentiation across scenarios for unsuccessful trials. When a pilot is successful, it may be related to adapting their monitoring strategy more to the challenge at hand as observed by a shift in focus on either the PFD or ND between challenges. This does not seem to be the case for unsuccessful trials for which the PDT distribution between the PFD and ND seems more similar between challenges.

Figure 7 depicts ANL for the 17 pilots in the same four challenges. Means per AoI across pilots are provided by red dashed lines and vertical lines provide the standard deviation for each individual on the AoI. Note that standard deviations are relatively large, indicating that Neglect Latencies varied considerably across pilots during the challenges. As pilots focused most on the PFD and ND (see Figure 6), ANLs for only those AoIs are presented here. Again, plots on the left provide results for successful pilots, while the plots on the right provide results for unsuccessful pilots. The top four plots provide ANL for the two challenges with waypoint restrictions (see also Appendix C). It can be observed that for S1C1, the Neglect Latency for the ND is lower than for the PFD. This does not seem to be the case for S4C2. Figure 7 indicates that overall ANL might be slightly higher for unsuccessful than successful pilots.

The bottom four plots of Figure 7 provide ANL for the two challenges concerning flap extensions. PFD and ND Neglect Latencies were similar to each other for both successful and unsuccessful pilots. A very low ANL or even values of zero can be observed for unsuccessful pilots of S2C4. Unsuccessful pilots moved the flap handle right after the PF's request to change flaps instead of waiting for the appropriate airspeed resulting in very short challenge durations. Finally, comparing all challenges might indicate again that successful pilots adapted their monitoring strategy more deliberately depending on the challenge. This trend in ANL is far less clear as compared to that in the PDT in Figure 6.

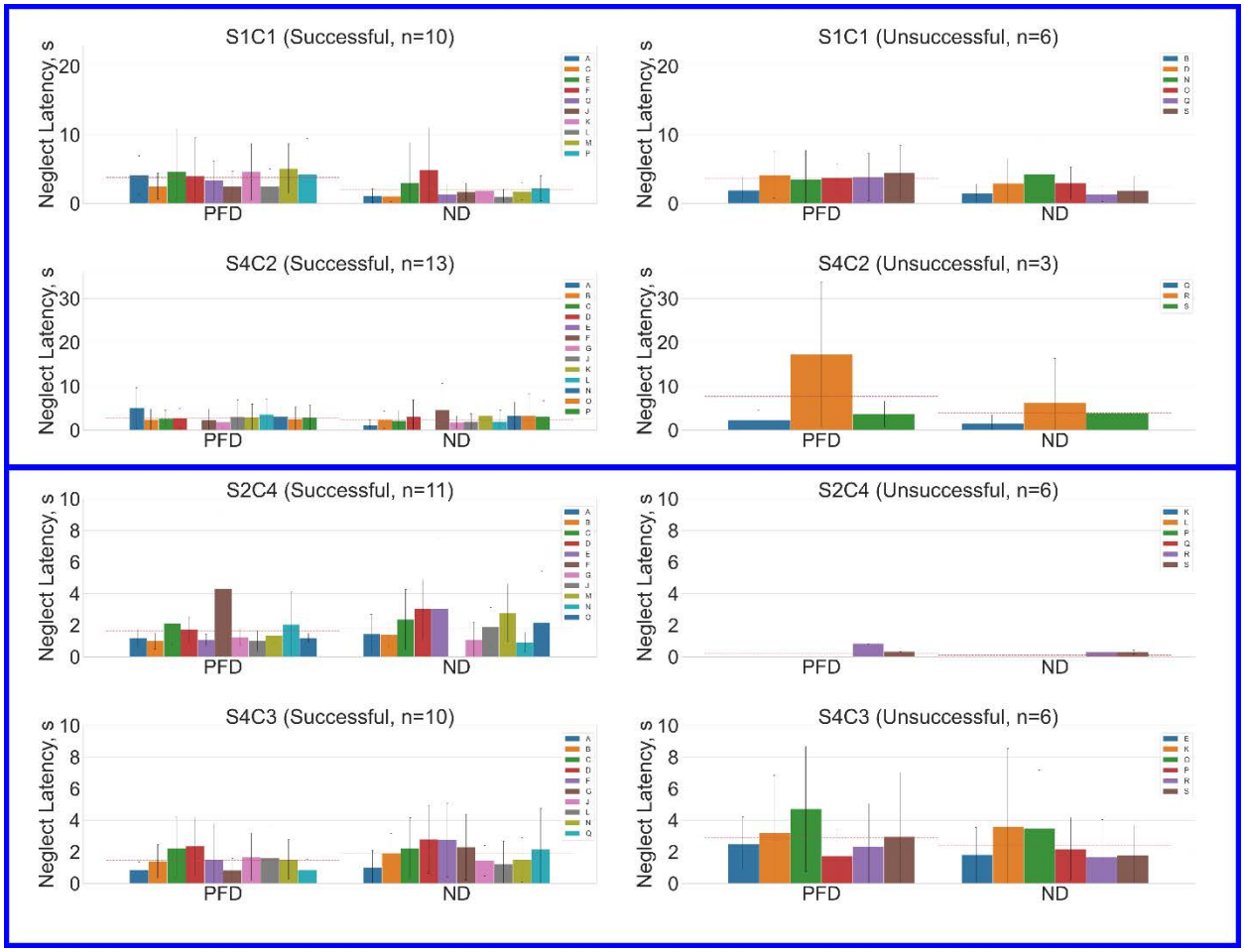

Figure 7. AoI Neglect Latency by individual participant (indicated by color) grouped by successful or unsuccessful outcome. Data for the four monitoring challenges where AOIs most relevant to success might be identified. S1C1 and S4C2 are challenges to meet waypoint restrictions; here altitude is critical to track and so the vertical situation display on the ND might be particularly helpful. S2C4 and S4C3 are challenges to extend the flaps at the appropriate time; here vertical speed is critical to track and so the VS display on the PDF might be particularly helpful.

3.8. Pilot Debriefing

Overall, pilots were quite positive about the experience. They evaluated training with five 7-point rating scales where high numbers are positive. Average ratings on each question were over 6 on each scale. No scores were on the negative side of a scale and only one score was neutral. Comparing ratings of the overall session to flying in the simulator, six participants rated the whole session greater than flying the simulator, while three rated the simulator higher than the overall session. Pilots were very positive about the value of training monitoring; had a variety of comments about communication in the context of monitoring; and had comments about what they liked (e.g., the interactives) and what could be improved (more interactives). Details about the debriefing results are in Appendix E.

4. Discussion

4.1. Importance and Goals of Research

Understanding and improving monitoring are important objectives for maintaining and improving aviation safety. We conducted a study with airline FOs acting as PM that tested the effect of our training intervention on monitoring. Further, it also tested for an effect on monitoring of FD on or off during approach. We characterize monitoring as an active process of making sense of the aircraft in the operational environment, particularly, the flight path. Effective monitoring is guided by a Situation Model, that is, the pilot's integrated understanding of the dynamic situation. Monitoring is a process of updating the Situation Model by selecting an important monitoring question; by finding relevant information about the current state or trajectory and comparing this with what is expected; and by identifying what actions might be needed. We developed a training intervention and realistic scenarios with behavioral markers to measure monitoring.

In addition to training, display configuration might impact monitoring. This study investigated whether flying a final approach with the FD turned off (and automation on), pilots might look around more widely, resulting in better monitoring.

4.2. Summary of Training Intervention Results

We provided training to 19 FOs of a major U.S. airline, acting as PM. Our training intervention presented monitoring as a sense-making activity, focusing on monitoring the flight path—where am I, where will I be, where is this in relation to my expectations and to requirements? The training emphasized monitoring as an active process of investigation. It introduced the concepts of a situation model, the cycle of updating the model, and of the importance of communicating to build a shared model. Examples focused on monitoring FPM, particularly management to meet ATC clearances. We sought active engagement of the pilot through questions and through brief exercises applying the monitoring concepts to scenarios or snippets that the pilot had flown in the immediately prior sim session, or in other experience.

The evaluation of training measured and compared performance before and after training. Monitoring performance was measured on challenges nested within the realistic scenarios. Each monitoring challenge presented an event that required a specific response by the PM, thus providing a behavior marker to measure monitoring effectiveness. We found modest but significant improvement in monitoring performance from the pre- to post-training. This suggests that monitoring can be improved by training and that training based on active sensemaking for monitoring FPM may be helpful.

4.3. Summary of Flight Director and Eye-Tracking Results

Note that pilots were not given any instructions about when or where to look at any time during the experiment. We assessed the hypothesis that turning the FD off during final approach would decrease time spent looking at the PFD on the two scenarios that continued to landing. During final approach there was a nonsignificant trend for the PMs to focus less on the PFD with the FD on (lower PDT) as compared to when the FD was off. This was a trend in the opposite direction of what was hypothesized; that is, the hypothesis was rejected. Nor were any clear effects of FD status found for the ANL.

Four challenge events were selected for detailed, exploratory examination of the eye-tracking data and possible relation to performance. Each of these four challenges had clear start and end times and they

had clearly relevant AoIs needed for success. Thus, we believed relations between performance and eye-tracking data might be most likely to be discovered here. In two challenges the pilots had to meet waypoint restrictions and in two the pilots had to extend flaps at the appropriate time. Pilots focused more on the PFD and ND compared to other AoIs in each of these four challenges. Inspection of Figures 6 and 7 suggests that pilots who successfully completed the challenges may have changed their attention allocation more across the different scenarios that unsuccessful pilots. Successful pilots may have focused more on the AoI that contained the most relevant information to successfully complete the challenge: ND for S1C1; ND and PFD for S4C2; and PFD for S2C4 and S4C3.

Our data did not find (nor was the study designed to discover) specific scan patterns or attention allocation contributing to success on individual monitoring challenges. Our data do suggest that successful pilots may allocate their gaze differently for different challenges, at least for challenges where performance depends heavily on a specific display. This suggests that part of successfully solving problems is to look for the information pertaining to those problems, instead of following a standard scan pattern, a suggestion in line with the implications of the sensemaking model of monitoring (Billman, et al., 2020, Mumaw et al., 2020).

4.4. Limitations and Future Research

Concerning training to improve pilots' monitoring, our study did produce results suggesting that our training and our measurement may be effective. However, there are several important limitations to the study. First, the test for impact of training was virtually immediate. Thus, the study does not speak to whether any impact would be retained for operationally relevant periods. Second, relatedly, while the flights flown and the details of the challenges changed from pre-training to post-training, pilots returned to the same setting and the same Captain. Possibly, this induced vigilance specific to this partner or setting. Third, performance change might be real and transferable but caused by participation in the pre-training simulator session. Possibly, addressing the challenges presented, either alone or in combination with training, might have been responsible for the change. This would be an interesting outcome as well but cannot be distinguished by the current study. Finally, we had a limited number of challenge events and of pilots. Replication with a new sample of events and of pilots would also be important.

Future research will need to replicate and extend this finding to make it of practical relevance and to assess the scope of retention over time, generalization to novel challenges, and transfer to more varied simulator settings. At least as important is development and extension of the training intervention itself. While this prototype provides an important start, more impactful training delivery can undoubtably be developed, incorporating a wide set of Instructional Design principles.

Concerning the use of eye-tracking, an increasing number of new flight simulators and training devices are equipped with eye-tracking technologies suggesting there will be an increasing interest in whether and how eye-tracking can be used effectively in training. More research would be required to determine whether and where there are patterns of attention distribution that contribute to success on specific types of situations and what eye-tracking measures are most appropriate to consider.

We only investigated the eye-tracking measures with respect to the main AoIs as defined in the eye-tracking software. In follow-on research it could be useful to look at specific areas on each AoI such as the speed or altitude tapes on the PFD. While this was technically possible with the data we collected, such AoI analysis would have required the pre-calibration of each individual pilot's eye gaze to ensure high precision eye-tracking data. Given this was not the core focus of this study, and

also due to time constraints, the Seeing Machines' calibration module was deliberately not supplied for the study to enable sub-AoI analysis. The eye tracker used in this study utilized one camera and two IR illuminators. In general, accurate results were obtained with respect to identifying when pilots were looking at the main AoIs, however, for two pilots the eye-tracking data were not sufficiently accurate. One of those pilots was wearing glasses, which can cause reflections that can disrupt the eye-tracking software. Issues like these can possibly be minimized with careful calibration for each individual pilot and mitigated with further optimization of the camera and infrared installation and orientation but will certainly need attention before eye tracking can truly be an integral part of pilots' training in flight simulators.

An important aspect of this study was simultaneous collection and analysis of the three types of data: video/audio; simulator; and eye-tracking data. We developed a custom application to integrate and synchronize these data using UTC time stamps. This was a nontrivial, important step to make sense of what pilots were doing and this allowed us to separate successfully from unsuccessfully managed events. Integrating multiple types of data is also an important aspect of utilizing eye tracking for future training and should receive adequate attention.

4.5. Conclusion

We characterize monitoring as the process of building and maintaining one's understanding of the unfolding situation. Doing this requires integrating general knowledge and expectations about how things work with the particulars of the specific event. Monitoring in this broad sense is a central part of the pilot's job. As automation increases, monitoring is likely to become an even bigger part of piloting and of aviation safety. Thus, maintaining and improving monitoring is an increasingly important problem. Design of interfaces, of procedures, and of training can all contribute.

For training, two complementary thrusts can drive improvement forward. As a broad approach, training monitoring as an active, prospective mode of cognition may help pilots be more engaged and find monitoring more interesting. As a set of specific techniques, training that provides context-specific strategies for what and how to monitor, given specific cues, may help pilots use their engagement and attention most productively. Together, these may improve training so that effective techniques do not need to be discovered through the experience of an individual pilot. If knowledge acquired through experience can be transmitted through training, this will accelerate acquisition of expertise.

Methods for assessing monitoring performance are an important complement to methods for training. Because monitoring is a heavily cognitive and internal task, it is difficult to assess. Nevertheless, operationally relevant, observable behavior can be used to measure monitoring. Two components are needed for effective assessment of monitoring. First, practices need to be established so that understanding the situation in a certain way means a particular action should be taken, either communication or some action on the airplane—if the PM understands that the aircraft cannot make the next waypoint, they should say something. These informative behaviors are often called behavioral markers. Second, the situations requiring a specific response need to be produced (as in the simulator) or recognized (as in observation of line flight). This approach will not track every nuance but will measure performance at the point it has operational impact. In some cases, finer-grained measurement may be important, for example, in assessing merits of alternative interface design.

Conversely, use of eye-tracking to measure or train monitoring remains challenging. In a complex environment the location, duration, and sequence of eye fixations is related to information extraction and use in complex ways, not well understood.

Much future work is needed to further investigate the complex, cognitive skills and knowledge needed for monitoring in complex dynamic environments. Our work provides an example of how monitoring by airline pilots can be measured and trained. We hope this study will provide a steppingstone for additional research that extends the current findings.

References

Billman, D, Mumaw, R., & Feary, M. (2020) A Model of Monitoring as Sensemaking: Application to Flight Path Management and Pilot Training. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 64). SAGE Publications.

Billman, D., Mumaw, R., Feary, M. (2020) *Methods for Evaluating the Effectiveness of Programs to Train Pilot Monitoring.* NASA/TM–20210000045.

Christiansen, B., 2019. Regression Models for Ordinal Data. R package version 2019. 12–10. https://CRAN.R-project.org/package=ordinal.

Bates D., Mächler M., Bolker B., Walker S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

FAA (2013). Operational use of flight path managements systems; Final report of the performance-based operations Aviation Rulemaking Committee / Commercial Aviation Safety Team Flight Deck Automation Working Group. Washington, DC: FAA.

ICAO (2016). Guidance material for improving flightcrew monitoring. Montreal, Quebec, Canada: ICAO.

Mumaw, R.J., Billman, D., & Feary, M. (2020). *Analysis of Pilot Monitoring Skills and a Review of Training Effectiveness*. NASA/TM-20210000047.

Salvucci, D.D., & Goldberg, J.H. (2000). Identifying fixations and saccades in eye-tracking protocols, *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '00*, ACM Press. https://doi.org/10.1145/355017.355028.

Appendix A. Simulator Variables Captured

Number	Variable	Description	Unit
1	frame_count	frame count	unit
2	sec_utc	UTC time in sec	Seconds
Automation state			
3	at_eng_mode	A/T Engaged Mode Capt and F/O	ASCII
4	capt_roll_eng	Roll Engaged Mode Capt	ASCII
5	capt_roll_arm	Roll Armed Mode Capt	ASCII
6	capt_pit_eng	Pitch Engaged Mode Capt	ASCII
7	capt_pit_arm	Pitch Armed Mode Capt	ASCII
8	capt_cws_pit	CWS P Engaged Mode Capt	0/1
9	capt_cws_roll	CWS R Engaged Mode Capt	0/1
10	capt_ap_stat	A/P Status Capt	ASCII
11	fo_roll_eng	Roll Engaged Mode F/O	ASCII
12	fo_roll_arm	Roll Armed Mode F/O	ASCII
13	fo_pit_eng	Pitch Engaged Mode F/O	ASCII
14	fo_pit_arm	Pitch Armed Mode F/O	ASCII
15	fo_cws_pit	CWS P Engaged Mode F/O	0/1
16	fo_cws_roll	CWS R Engaged Mode F/O	0/1
17	fo_ap_stat	A/P Status F/O	ASCII
18	capt_cmd_pit_dev	Pitch CMD Dev #1	DOT
19	fo_cmd_pit_dev	Pitch CMD Dev #2	DOT
20	capt_cmd_roll_dev	Roll CMD Dev #1	DOT
21	fo_cmd_roll_dev	Roll CMD Dev #2	DOT
22	ils_1_gs_dev	ILS #1 G/S Dev	DOT
23	ils_1_loc_dev	ILS #1 Loc Dev	DOT
24	ils_2_gs_dev	ILS #2 G/S Dev	DOT
25	ils_2_loc_dev	ILS #2 Loc Dev	DOT
MCP Displays			
26	mcp_crs_1_ds	Course #1	Degrees
27	mcp_crs_2_ds	Course #2	Degrees
28	mcp_ias_mach_ds	IAS/MACH Flashing - 'Axxx' - under speed and 8'xxx' - overspeed	knot/mach
29	mcp_hdg_ds	Heading	Degrees
30	mcp_alt_ds	Altitude	feet
31	mcp_vert_spd_ds	Vert Speed	feet/second
MCP Buttons/led			
32	mcp_n1	N1	Note 1
33	mcp_spd	Speed	Note 1
34	mcp_lvl_spd	Lvl Spd	Note 1
35	mcp_vnav	VNAV	Note 1
36	mcp_lnav	LNAV	Note 1
37	mcp_vor_loc	VOR/LOC	Note 1
38	mcp_apprh	APP	Note 1

39	mcp_hdg_sel	Hdg Sel	Note 1
40	mcp_alt_hld	Alt Hold	Note 1
41	mcp_vert_spd	V/S	Note 1
42	mcp_cmd_a	CMD A	Note 1
43	mcp_cmd_b	CMD B	Note 1
44	mcp_cws_a	CWS A	Note 1
45	mcp_cws_b	CWS B	Note 1
MCP Switches			
46	mcp_fd_1	F/D #1	0/1
47	mcp_fd_2	F/D #2	0/1
48	mcp_at_arm	A/T Arm	0/1
49	mcp_cross_over	C/O	0/1
50	mcp_spd_intv	Spd Intv	0/1
51	mcp_bank_angle_max	Bank Angle MAX (10, 15, 20, 25, 30)	10,15,20,25,30
52	mcp_alt_intv	Alt Intv	0/1
MCP Lts			
53	mcp_mstr_lt_1	Master #1	0/1
54	mcp_mstr_lt_2	Master #2	0/1
55	mcp_at_arm_lt	A/T Arm	0/1
Capt Efis control			
56	cap_ef_min_alt	Min Alt (Radio, baro) 1=radio 2=baro 0 = dead bus	Note 2
57	cap_ef_min_adj	Adj Min 2,1,0,-1,-2	Note 3
58	cap_ef_min_cur_alt	Minimum Descent Altitude(baro), or Decision Height(radio)	feet
59	cap_ef_min_rst	MINS RST	0/1
60	cap_ef_fpv	FPV display	Note 1
61	cap_ef_mtrs	Mtrs display	Note 1
62	cap_ef_baro	Baro (IN, HPA) 1=IN 2=HPA 0=dead bus	Note 4
63	cap_ef_baro_adj	Adj Baro 2,1,0,-1,-2	Note 3
64	cap_ef_baro_cur	Baro current (MB)	MB
65	cap_ef_baro_std	BARO STD	Note 1
66	cap_ef_1_ptr	#1 pointer (vor, none, adf) 0-none, 1-VOR, 2-ADF	Note 5
67	cap_ef_nd_mode	ND Mode (App,VOR,MAP,PLN) 1 - app 2 - vor 3 - map 4 - plan 0 - dead bus	Note 6
68	cap_ef_ctr_nd	CTR ND	Note 1
69	cap_ef_rnge	Range (5,10,20,40,80,160,320,640)	NM
70	cap_ef_tfc	TFC	Note 1
71	cap_ef_2_ptr	#2 pointer (vor, none ,adf) 0-none, 1-VOR, 2-ADF	Note 5
72	cap_ef_wxr	wxr display	Note 1
73	cap_ef_sta	sta display	Note 1
74	cap_ef_wpt	wpt display	Note 1
75	cap_ef_arpt	arpt display	Note 1

76	cap_ef_data	data display	Note 1
77	cap_ef_pos	pos display	Note 1
78	cap_ef_terri	ter display	Note 1
		F/O Efis control	
79	fo_ef_min_alt	Min Alt (Radio, baro) 1=radio 2=baro 0 = dead bus	Note 2
80	fo_ef_min_adj	Adj Min 2,1,0,-1,-2	Note 3
81	fo_ef_min_cur_alt	Minimum Descent Altitude(baro), or Decision Height(radio)	feet
82	fo_ef_min_rst	MINS RST	0/1
83	fo_ef_fpv	FPV display	Note 1
84	fo_ef_mtrs	Mtrs display	Note 1
85	fo_ef_baro	Baro (IN, HPA) 1=IN 2=HPA 0=dead bus	Note 4
86	fo_ef_baro_adj	Adj Baro 2,1,0,-1,-2	Note 3
87	fo_ef_baro_cur	Baro current (MB)	MB
88	fo_ef_baro_std	BARO STD	Note 1
89	fo_ef_1_ptr	#1 pointer (vor, none, adf) 0-none, 1-VOR, 2-ADF	Note 5
90	fo_ef_nd_mode	ND Mode (App,VOR,MAP,PLN) 1 - app 2 - vor 3 - map 4 - plan 0 - dead bus	Note 6
91	fo_ef_ctr_nd	CTR ND	Note 1
92	fo_ef_rnge	Range (5,10,20,40,80,160,320,640)	NM
93	fo_ef_tfc	TFC	Note 1
94	fo_ef_2_ptr	#2 pointer (vor, none ,adf) 0-none, 1-VOR, 2-ADF	Note 5
95	fo_ef_wxr	wxr display	Note 1
96	fo_ef_sta	sta display	Note 1
97	fo_ef_wpt	wpt display	Note 1
98	fo_ef_arpt	arpt display	Note 1
99	fo_ef_data	data display	Note 1
100	fo_ef_pos	pos display	Note 1
101	fo_ef_terri	ter display	Note 1
		Aircraft variables	
102	cal_as	Calibrated airspeed	Knots
103	true_as	True airspeed	Knots
104	gnd_spd	Ground speed	Knots
105	rate_of_clb	Rate of climb	Feet/minute
106	pres_alt	Pressure altitude	Feet
107	height_above_teri	Height above ground	Feet
108	radio_alt	Radio altitude	Feet
109	pitch_angle	Pitch angle	Degrees
110	roll_angle	Roll angle	Degrees
111	hdg_angle	Heading angle	Degrees
112	angle_of_attack	Angle of attack	Degrees
113	sideslip_angle	Sideslip angle	Degrees

114	loc_dev	Localizer deviation in dots	Dot
115	glideslope_dev	Glideslope deviation in dots	Dot
116	lat	Aircraft latitude	Degrees
117	long	Aircraft longitude	Degrees
118	on_gnd	On-ground flag	0/1
119	mag_track_angle	Magnetic track angle	Degrees
120	mag_hdg_angle	Magnetic heading angle	Degrees
121	x_pos_rwy_td	x-position to touchdown	Feet
122	y_pos_rwy_td	y-position to touchdown	Feet
123	z_pos_rwy_td	z-position to touchdown	Feet
124	column_pos	Column position (CP+FO)	Degrees
125	wheel_pos	Wheel position (CP+FO)	Degrees
126	rudder_pos	Rudder pedal position (CP+FO)	Degrees
127	rudder_trim	Rudder trim setting	Degrees
Throttle lever angle (left+right)			
128	throttle_1_pos	#1	Degrees
129	throttle_2_pos	#2	Degrees
Aileron position (left+right)			
130	ailerion_l_pos	Left	Degrees
131	aileron_r_pos	Right	Degrees
Elevator position (left+right)			
132	elevator_l_pos	Left	Degrees
133	elevator_r_pos	Right	Degrees
134	rudder_pos	Rudder position	Degrees
135	hori_stab_pos	Horizontal stabilizer position	Degrees
136	flap_angle	Flap angles	Degrees
137	spoiler_pos	Spoiler positions	Degrees
138	pitch_rate	Pitch rate	deg/sec
139	roll_rate	Roll rate	deg/sec
140	yaw_rate	Yaw rate	deg/sec
Body acceleration (x,y,z)			
141	x_body_accel	x	ft/s*2
142	y_body_accel	y	ft/s*2
143	z_body_accel	z	ft/s*2
Load factor			
144	load_fac_x	Load Factor at CG X-Body	g
145	load_fac_y	Load Factor at CG Y-Body	g
146	load_fac_z	Load Factor at CG Z-Body	g
147	static_temp	Static Temperature	deg C
N1 (engine 1+2)			
148	eng_n1_1	1	%
149	eng_n1_2	2	%
Fuel lever (left+right)			
150	eng_fuel_lvr_1	#1	0/1
151	eng_fuel_lvr_2	#2	0/1

Fuel tank (left+center+right)			
152	l_fuel_tank_qty	Left Qty	lbs
153	ctr_fuel_tank_qty	Center Qty	lbs
154	r_fuel_tank_qty	Right Qty	lbs
Turbulence components			
155	turb_x_comp	x	ft/sec
156	turb_y_comp	y	ft/sec
157	turb_z_comp	z	ft/sec
Wind components			
158	wind_dir_at_ac	wind direction at aircraft	Degrees
159	wind_spd_at_ac	wind speed at aircraft	Knots
Stick shaker			
160	stick_shaker_cap	Capt	0/1
161	stick_shaker_fo	F/O	0/1
Brake pedal position (left+right)			
162	brake_pos_l	Left	Degrees
163	brake_pos_r	Right	Degrees
Miscellaneous			
164	mstr_caution	Master Caution Light capt side	Note 1
165	capt ap_discon	Capt A/P disconnect switch	0/1
166	fo_ap_discon	F/O A/P disconnect switch	0/1
167	ap_caut_lt	AP Caution Light	0/1
168	ap_warn_lt	AP Warning Light	0/1
169	ap_discon_horn	A/P disconnect horn	0/1
170	alt_warn_horn	Altitude Warning Audio	0/1
171	ovr_spd_clkr	Overspeed Clacker Audio	0/1
172	at_1_discon	#1 A/T disconnect switch	0/1
173	at_2_discon	#2 A/T disconnect switch	0/1
174	at_caut_lt	A/T Caution Light	0/1
175	at_warn_lt	A/T Warning Light	0/1
176	to_sw_1	#1 T/O switch	0/1
177	to_sw_2	#2 T/O switch	0/1
178	to_warn_lt	T/O warning Light	0/1
179	lnding_gear_handle	Landing gear handle (Up, Off, Down)	Note 7
180	l_lnd_gr_pos_lt	Left Landing gear pos lts	Note 9
181	n_lnd_gr_pos_lt	Center Landing Gear pos lts	Note 9
182	r_lnd_gr_pos_lt	Right Landing Gear pos lts	Note 9
183	tiller_pos	Tiller Position	Degrees
184	cap_trim_sw	Capts stab trim switches (UP,None,Down)	Note 8
185	fo_trim_sw	F/O stab trim Switches (Up,none, Down)	Note 8
186	st_trim_main_off	Stab trim Main override switch	0/1
187	st_trim_ap_off	Stab trim Autopilot cutoff switch	0/1
188	FMC_alert_lt	FMC ALERT Caution Light	0/1

189	V1	FMC V1	knots
190	VR	FMC VR	knots
191	v2	FMC V2	knots
192	vref	FMC Vref	knots
193	capt_vsd_on	Capt displaying VSD	0/1
194	fo_vsd_on	F/O displaying VSD	0/1
195	auto_bk_sw	Auto Brake Switch Pos	Note 10
196	auto_bk_inop_lt	Auto Brake Inop Light	0/1
197	anti_skid_inop_lt	Anti Skid Inop Light	0/1
198	prk_brk_lt	Parking Brake Light	0/1
199	event_mkr	Event Marker	0/1
200	spd_brk_arm	speed brake armed Light	0/1
201	spd_brk_do_not_arm	speed brake do not arm Light	0/1
202	to_horn	takeoff horn	0/1
203	ail_trim_r	Aileron trim	degrees
204	rnp_vert	RNP for vert	feet
205	anp_vert	ANP for vert	feet
206	rnp_lat	rnp for lat	
207	anp_lat	anp for lat	NM
208	rnp_vert_dev	rnp vert dev	feet
Miscellaneous cockpit lights			
209	stab_out_trim	Stab out of trim Light	0/1
210	cabin_alt_wrn_lt	Cabin altitude warning Light	0/1
211	capt_bel_gs_lt	capt below glideslope	Note 1
212	fo_bel_gs_lt	f/o below glideslope	Note 1
213	le_flap_tran_lt	LE Flaps Transit Light	0/1
214	LE_flap_ext_lt	LE Flap Ext	0/1
215	mc_capt_flt_contrl	mc capt side Flt Contrl	0/1
216	mc_capt_irs	mc capt side IRS	0/1
217	mc_capt_fuel	mc capt side Fuel	0/1
218	mc_capt_elec	mc capt side Elec	0/1
219	mc_capt_apu	mc capt side APU	0/1
220	mc_capt_det	mc capt side Ovht/Dec	0/1
221	mc_cap_recall	mc capt side recall pushed	0/1
222	mc_fo_anti_ice	mc fo side anti-ice	0/1
223	mc_fo_hyd	mc fo side hyd	0/1
224	mc_fo_doors	mc fo side doors	0/1
225	mc_fo_eng	mc fo side eng	0/1
226	mc_fo_ovrhd	mc fo side overhead	0/1
227	mc_fo_aircond	mc fo side aircond	0/1
228	mc_fo_recall	mc fo side recall pushed	0/1
229	fire_wrn_lts	capt or fo fire warning	Note 1
230	fo_mc_lts	fo master cautiion Light	Note 1
231	spdbrk_ext_lt	Speedbrake EXT Light	0/1

Notes:	
1	0 = no light or mode and no button pushed
	1 = light or mode set and no button pushed

	2 = no light or mode and button pushed
	3 = light or mode set and button pushed
2	0 = dead arinc bus
	1 = radio selected
	2 = baro selected
3	2 = fast increase
	1 = slow increase
	0 = no adjust
	-1 = slow decrease
	-2 = fast decrease
4	0 = dead arinc bus
	1 = IN selected
	2 = HPA selected
5	0 = none selected
	1 = Vor selected
	2 = adf selected
6	0 = dead arinc bus
	1 = app selected
	2 = vor selected
	3 - map selected
	4 = plan selected
7	-1 = landing gear handle down
	0 = landing gear handle off
	1 = landong gear handle up
8	-1 = nose down
	0 = no switch actived
	1 = nose up
9	0 = both light off
	1 = gear down lts (green) on and gear up lts (red) off
	2 = gear down lts (green) off and gear up lts (red) on
	3 = gear down lt (green) on and gear up lt (red) on
10	0 = autobrake set off
	1 = autobrake set to 1
	2 = autobrake set to 2
	3 = autobrake set to 3
	4 = autobrake set to MAX
	5 = autobrake set to RTO

Appendix B. Eye-Tracking Variables Captured

Number	Variable	Description	Unit
1	DAS_Conf	Confidence/error of eye tracking data (estimation)	0-1
2	DAS Quality	Indicates if the eyes are visible (estimation)	EYES VISIBLE/EYES NOT VISIBLE
3	DAS Object ID	ID of the AOI looked at by the pilot (estimation)	Note 1
4	DAS Object Name	AOI name (estimation)	Note 1
5	PGaze Object ID	AOI name (raw data)	Note 1
6	PGaze X Intersection	x coordinate of gaze on AOI (raw data)	0-1
7	PGaze Y Intersection	y coordinate of gaze on AOI (raw data)	0-1
8	PGaze Valid	Availability of eye gaze data (raw data)	TRUE/FALSE
9	Gaze Confidence	Confidence/error of eye tracking data	0-1
10	Gaze RMSE	Estimate of 1-sigma error	rad
11	Unified Gaze Source	Indicates the source method of derivation of the unified gaze ray	Note 2
12	Eyes Trackable	Indicates if the eyes are trackable	TRUE/FALSE
13	Eyewear Detected	Indicates if eyewear was detected	TRUE/FALSE
14	UTC Time (Microseconds)	UTC time in microseconds	microseconds
15	UTC Time (Seconds)	UTC time in seconds	Seconds

Notes:	
1	0 = Uknown
	1003 = PFD
	1004 = ND
	1005 = OTW
	1006 = Upper EICAS
	1007 = EFIS Control
	1008 = FMS
	1009 = MCP
	1010 = Lower EICAS
2	INVALID
	BOTH EYES TRACKING
	ONE EYE TRACKING
	HEAD TRACKING

Appendix C. Description of the Four Scenarios, Each Challenge, and Performance Scoring

Four scenarios were developed, each with three or four monitoring challenges. Scenarios 1 and 2 were each a descent and arrival into Dulles airport (KIAD). Scenarios 3 and 4 were each a descent and arrival into Las Vegas airport (KLAS). All scenarios began just prior to the top of descent (T/D) point and ended with either the initiation of a go-around or a landing at the airport. Scenarios 1 and 3 had a parallel structure; scenarios 2 and 4 were also parallel in structure. Each scenario operated in Instrument Meteorological Conditions to an altitude of approximately 500' AGL.

We created a four-point scoring scheme for each monitoring challenge; two outcomes were considered acceptable (Success: 4; Less than ideal: 3) and two outcomes were considered to be failures (Undesirable: 2; Bad: 1). In a few cases, one or two of these four outcomes were undefined because they were not operationally meaningful; these are coded as NA in the detailed description in this appendix and in Table 6 in the main body of this report.

After scoring, there was concern that for two events—Event 1 (first challenge in Scenario 1) and Event 9 (first event in Scenario 3)—performance initially coded as 2 was more appropriately coded as 3. This changed five events from a score of 2 to 3. The key analysis was redone with this scoring with very similar outcome. The clmm model statistics for training were $\chi2(2) = 11.337$, $p = .00345$; for order, $\chi2(2) = 2.117$, $p = .347$, and for training X order $\chi2(1) = .556$, $p = .456$. All results in the body of this report used the scoring reported below, not the modification described in this paragraph.

At a high-level, these were the monitoring challenges (events) in each scenario:

Scenario 1
1. Ev1 High on path- slowed by ATC
2. Ev2 Inappropriate mode (VNAV) – PF does not change
3. Ev3 Wrong altimeter setting
4. Ev4 Failure to set field elevation

Scenario 2
1. Ev5 Inappropriate mode (VS) – interaction of autoflight and PF action
2. Ev6 Shortened lateral path – ATC gives direct-to
3. Ev7 Inappropriate mode (LNAV) – PF selection
4. Ev8 Airspeed error – PF calls flaps when to fast

Scenario 3
1. Ev9 High on path - held by ATC
2. Ev10 Inappropriate mode (HDG SEL) – PF selection
3. Ev11 False glideslope

Scenario 4
1. Ev12 Inappropriate mode (LVL CHG) – PF selection
2. Ev 13 Shortened lateral path – ATC gives direct-to
3. Ev14 Airspeed error – PF fails to call for flaps 5
4. Ev15 Inappropriate mode (fails to arm APP)

Detailed descriptions of these monitoring challenges (events) are presented next.

Scenario 1
FLIGHT PLAN (KSEA – KIAD); MGW.GIBBZ3.IAD; ILS 01L

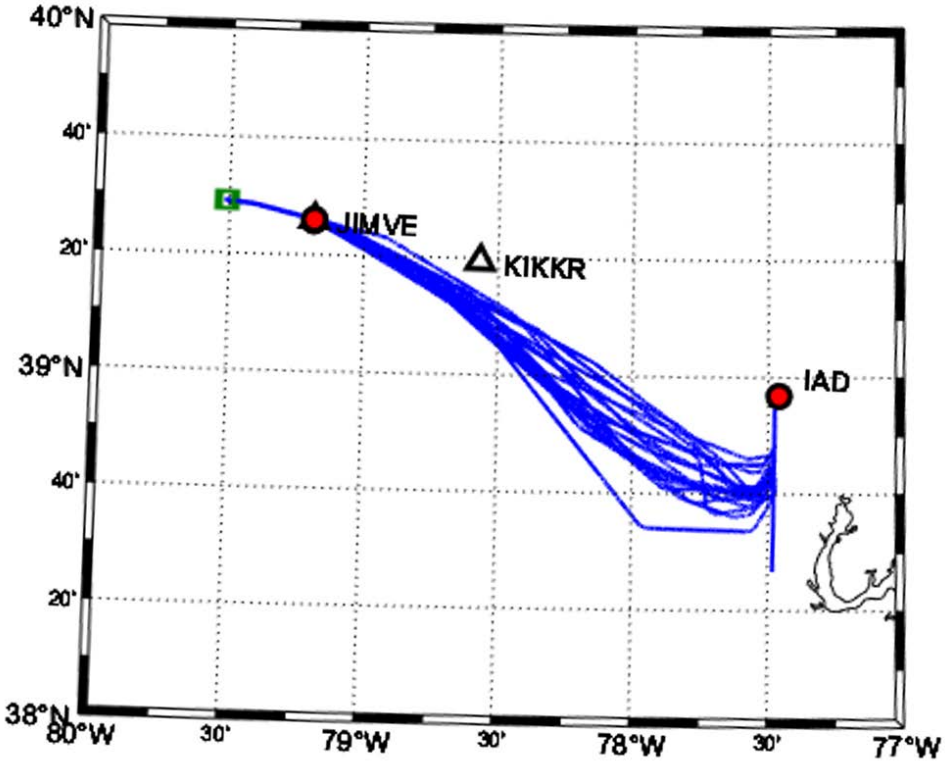

Figure C1. Map view of Scenario 1 with the flight track for all pilots. Green squares are the top of descent (start of the run) or the airport (end of the run). White triangles represent the waypoints and red dots indicate important events in challenges such as crossing a waypoint.

Challenge 1 (high on path): At T/D the flightcrew was given this clearance: "*AirlineName* 1, descend via the GIBBZ 3 RNAV Arrival, except maintain 250 knots." This clearance, due to an assigned airspeed lower than planned, forced the airplane above the FMS-derived flight path. There was a waypoint altitude restriction at JIMVE, which would be nearly impossible to meet with the lower airspeed. The following performance outcomes were defined:

1. Success: PM comments on not making the restriction at JIMVE or needing to descend more steeply; PM asks ATC for relief from the restriction at JIMVE.
2. Less than ideal: PM asks for relief.
3. Undesirable: PM only mentions that the airplane is going to be high at JIMVE; doesn't ask for relief.
4. Bad: PM does not mention or make any effort to address being high before JIMVE; shows a total lack of awareness.

Challenge 2 (inappropriate mode): After passing JIMVE, ATC requests "*AirlineName* 1, turn right heading 130, descend and maintain 11,000'." In this case, the PF remains in VNAV although the airplane is off the LNAV path and cleared below the next flight plan restriction, which is 14,000' at

KIKKR. If the airplane remains in VNAV, it will level off at KIKKR, which is not the clearance. The following performance outcomes were defined:

1. Success: PM sees that it is still in VNAV, realizes that there is a restriction at KIKKR, requests transitions to LVL CHG or VS prior to beginning to level at KIKKR.
2. Less than ideal: Airplane starts to level at KIKKR (14,000') but the PM quickly verbalizes the problem and prompts PF to continue descending.
3. Undesirable: Airplane levels at KIKKR (14,000').
4. Bad: Airplane levels at 14,000' for more than 10 seconds.

Challenge 3 (wrong altimeter setting): The flightcrew is given a bad altimeter setting, which will cause them to be 500' low on the approach. There is an aural alert that occurs around 5500' msl that says “Altimeter setting.” However, there is no guidance for this message in the flightdeck. They will need to perform a go-around. The following performance outcomes were defined:

1. Success: PM inquires to ATC about altimeter (altimeter setting, altitude), identifies that something is wrong and that there is a need for a go-around prior to breaking out of the clouds (around 500').
2. Less than ideal: PM determines there is a problem and requests go-around after breaking out of clouds.
3. Undesirable: PM requests a go-around when the EGPWS aurals occur.
4. Bad: PF initiates a go-around or enquires to PM about what to do or continues approach.

Challenge 4 (failure to set field elevation): After being cleared for the approach, it is appropriate to set the MCP altitude to the field elevation on an RNAV approach. In this case, the PF intentionally fails to set the field elevation, leaving the last cleared altitude (1900'; 3000' prior to that) in the MCP altitude window. If this altitude is not set lower, the airplane will capture and level at 1900, ' interrupting the approach. The following performance outcomes were defined:

1. Success: PM sets MCP altitude to field elevation prior to any problems with levelling off.
2. Less than ideal: Airplane levels at last cleared altitude (1900' or 3000') and goes into ALT HOLD briefly but immediately regains VNAV.
3. Undesirable: PM levels at last cleared altitude (1900' or 3000') and goes into ALT HOLD; and PF then has to use another vertical mode to get back to the approach path.
4. Bad: The flightcrew is unable to complete the approach because they did not set field elevation. Note: For purposes of this study, PF did not let it get this far, so this was not a possible outcome.

Scenario 2
FLIGHT PLAN (KBOS – KIAD); BAF.HYPER7.IAD; RNAV Y 19L

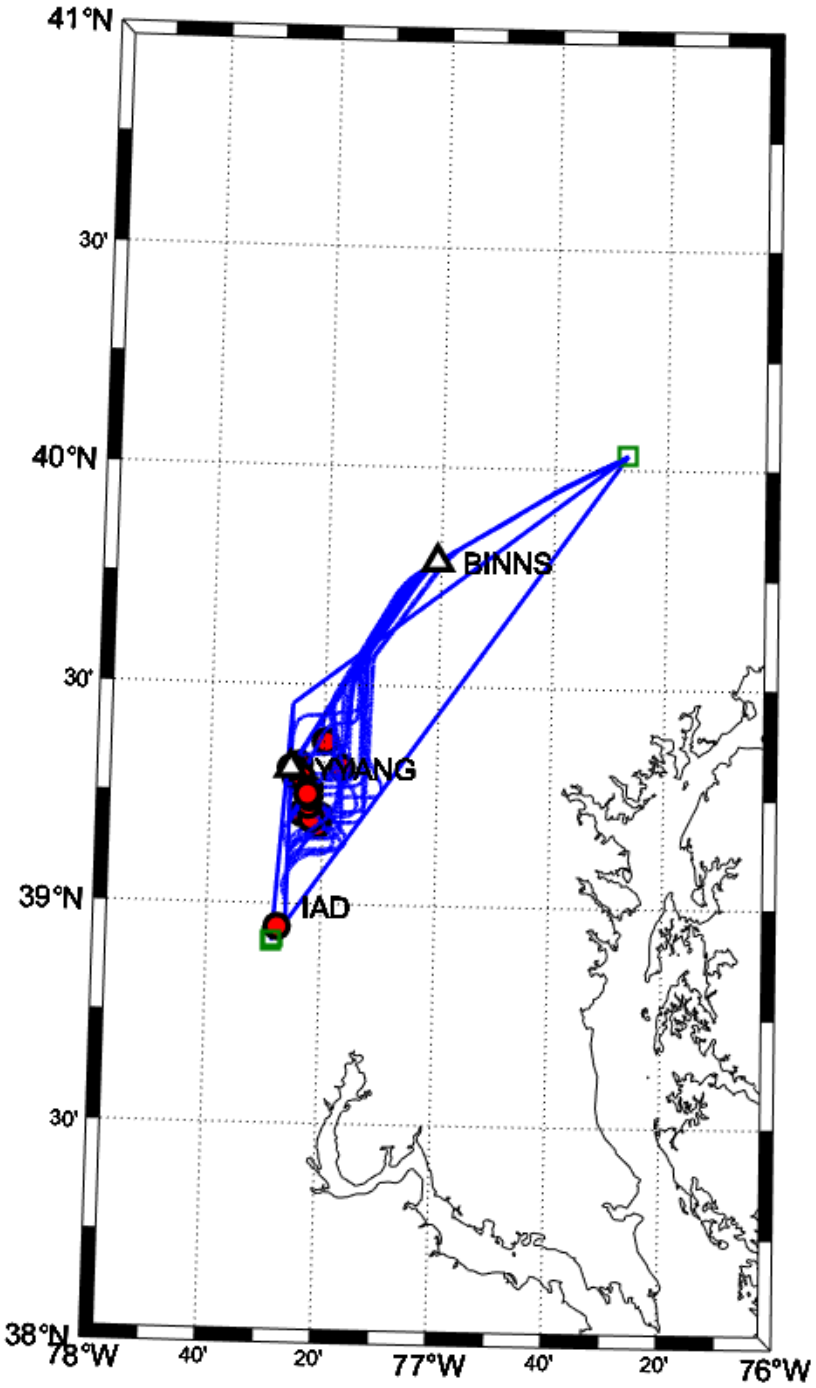

Figure C2. Map view of Scenario 2 with the track for all pilots. The straight line between top of descent and IAD is a repositioning of the simulator. Green squares are the top of descent (start of the run) or the airport (end of the run). White triangles represent the waypoints and red dots indicate important events in challenges such as crossing a waypoint.

Challenge 1 (inappropriate mode): The airplane is initially given an altitude clearance of FL180. Just prior to reaching FL180, ATC issues clearance to descend via the arrival ("*AirlineName* 1,

descend via the Hyper 7 RNAV Arrival."). The PF begins dialing down the MCP altitude while the FMA displays ALT ACQ, and for this airplane, this action causes a reversion into vertical speed (VS) mode. In VS mode, the airplane will not manage an altitude restriction at LIRCH appropriately. The following performance outcomes were defined:

1. Success: PM identifies and verbalizes the transition to VS and directs the PF back to VNAV or manages the restriction at LIRCH in VS.
2. Less than ideal: NA.
3. Undesirable: PM fails to see VS mode but aircraft makes the altitude restriction at LIRCH.
4. Bad: PM fails to see VS mode and aircraft misses the altitude restriction at LIRCH (altitude bust).

Challenge 2 (shortened lateral path): ATC clears the airplane direct to YYANG: "*AirlineName* 1, cleared direct YYANG, Cross YYANG at 3000'." This reduces the number of track miles flown but required the same altitude loss. It can be hard to descend and slow down with the shorter route. The following performance outcomes were defined:

1. Success: PM comments on not making it or needing to descend more steeply; asking for relief.
2. Less than ideal: NA.
3. Undesirable: NA.
4. Bad: Airplane fails to cross YYANG at 3000'; crosses high without ATC permission.

Challenge 3 (inappropriate mode): During ATC vectors to connect to the approach, the PF intentionally (and inappropriately) selects LNAV, which will intercept the track on the heading when LNAV was selected instead of flying the ATC-assigned vector to the approach (in this case approximately 80-degree intercept angle instead of approximately 30-degree intercept angle). The following performance outcomes were defined:

1. Success: PM identifies LNAV mode immediately (when free from other duties) and requests transition back to HDG SEL.
2. Less than ideal: NA.
3. Undesirable: NA.
4. Bad: PM fails to request transition to HDG SEL prior to airplane turning onto the LNAV track.

Challenge 4 (airspeed error): As the airplane is slowing for the approach, the PF calls for flaps 25 when the airspeed is around 190, which is too fast to deploy flaps 25. Ideally, the PM is aware that speed is too high and waits. The following performance outcomes were defined:

1. Success: when PF calls for Flaps 25, PM checks speed, verbalizes that speed is high and waits until airspeed is appropriate for Flaps 25.
2. Less than ideal: when PF calls for Flaps 25, PM waits until airspeed is appropriate for Flaps 25.
3. Undesirable: PM selects Flaps 25 but moves handle back to Flaps 15 after he/she realizes it was too soon.
4. Bad: PM selects Flaps 25 prior to airspeed getting low enough.

Scenario 3
FLIGHT PLAN (KSEA – KLAS); BTY.SUNST4.LAS; ILS 26L

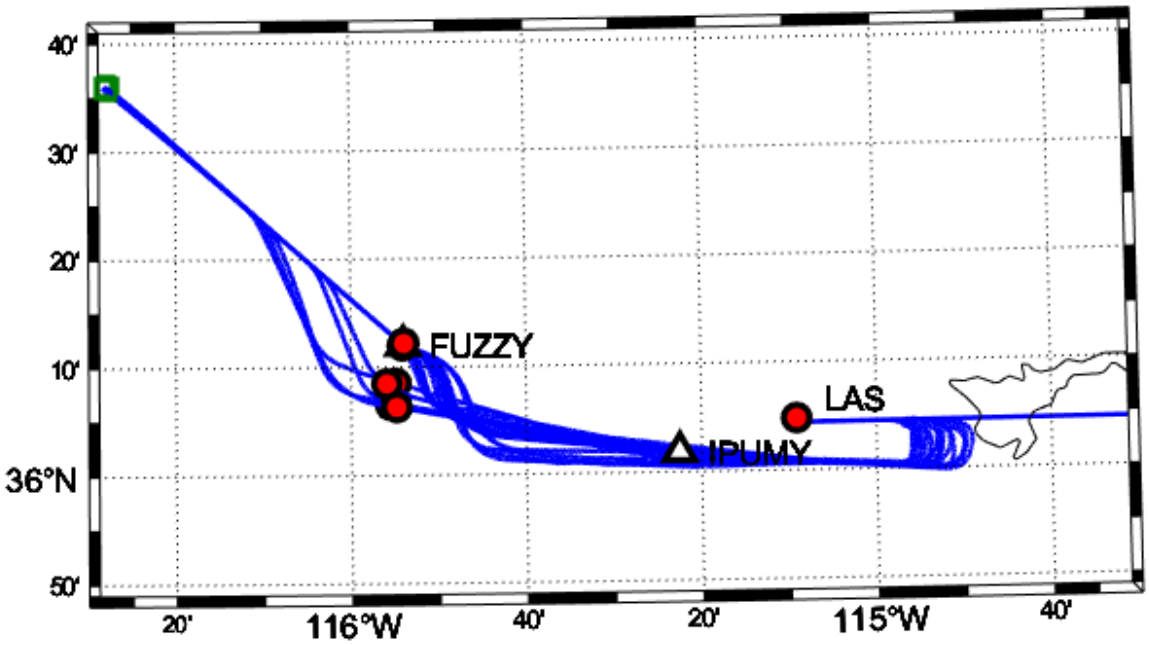

Figure C3. Map view of Scenario 3 with the track for all pilots. Green squares are the top of descent (start of the run) or the airport (end of the run). White triangles represent the waypoints and red dots indicate important events in challenges such as crossing a waypoint.

Challenge 1 (high on path): The airplane is held at cruise altitude past the T/D point and is eventually cleared to descend. This situation makes it very difficult to make the altitude restriction at FUZZY. The following performance outcomes were defined:

1. Success: PM comments on not making the restriction at FUZZY or needing to descend more steeply; PM asks ATC for relief from the restriction at FUZZY.
2. Less than ideal: PM asks for relief.
3. Undesirable: PM only mentions that the airplane is going to be high at FUZZY; doesn't ask for relief.
4. Bad: PM does not mention or make any effort to address being high before FUZZY; shows a total lack of awareness.

Challenge 2 (inappropriate mode): ATC clears them direct to IPUMY: "*AirlineName* 1, cleared direct to IPUMY, descend via the SUNST 4 RNAV Arrival." Because they are cleared to descend via the arrival, they need to stay in LNAV and VNAV. The PF intentionally (and inappropriately) selects HDG SEL as the lateral mode. Eventually, the airplane's track diverges from the LNAV flight path. The following performance outcomes were defined:

1. Success: PM comments that mode is still in HDG SEL instead of LNAV.
2. Less than ideal: NA
3. Undesirable: PM lets the mode remain in HDG SEL.
4. Bad: PM lets the mode remain in HDG SEL until the course deviation equals 1.0 mile.

Challenge 3 (false G/S): A false glideslope (G/S) is used to guide the airplane's vertical approach path. This false G/S is steeper than the true G/S, and the airplane descends below the true G/S. The following performance outcomes were defined:

1. Success: PM identifies that something is wrong and that there is a need for a go-around prior to breaking out of the clouds.
2 Less than ideal: PM determines there is a problem and requests a go-around after breaking out of clouds.
3. Undesirable: PM requests a go-around when the EGPWS aurals occur.
4. Bad: PF initiates a go-around or enquires to PM about what to do or continues approach.

Scenario 4
FLIGHT PLAN (KJFK – KLAS); LRAIN.TYSSN5.LAS; ILS 26L

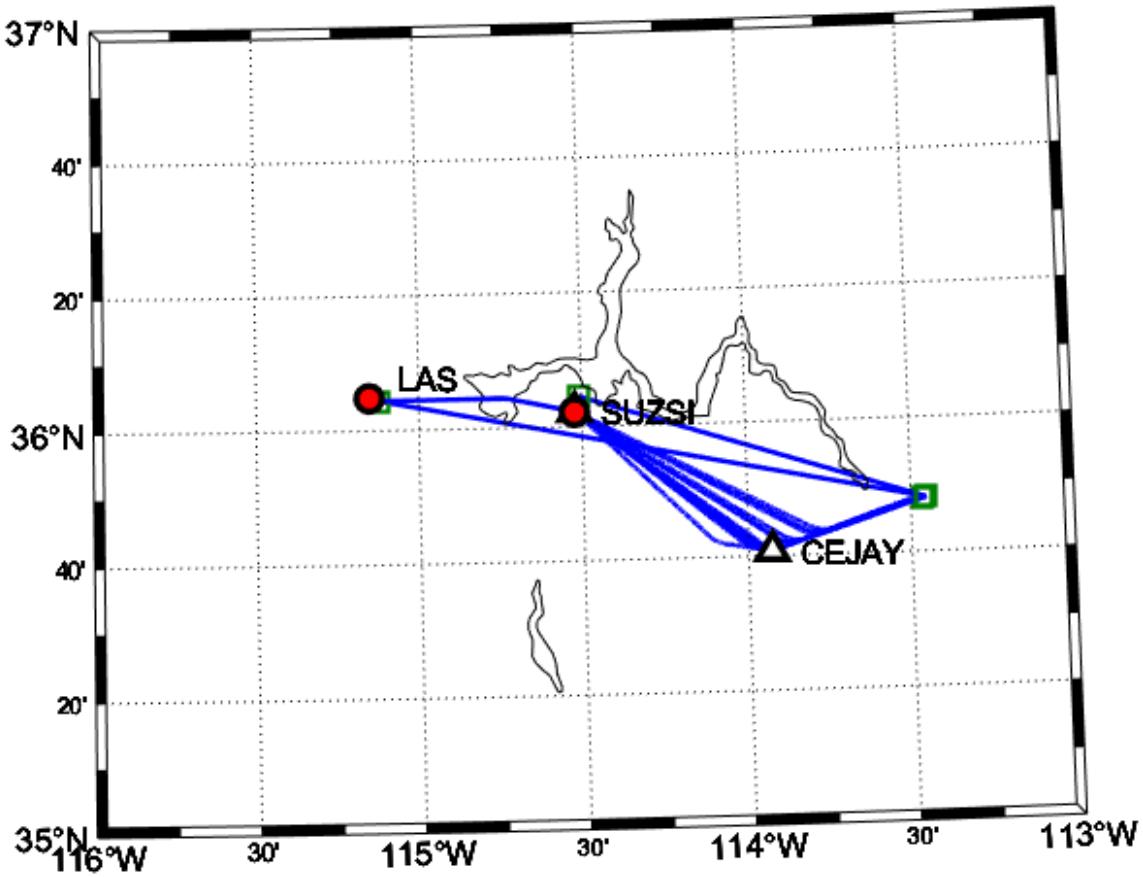

Figure C4. Map view of scenario 4 with the track for all pilots. Green squares are the top of descent (start of the run) or the airport (end of the run). White triangles represent the waypoints and red dots indicate important events in challenges such as crossing a waypoint.

Challenge 1 (inappropriate mode): Just after T/D, ATC gives this clearance: "*AirlineName* 1, descend via the TYSSN 5 RNAV arrival except maintain 250 knots." The requirement to descend via the arrival means the flightcrew should VNAV and LNAV, especially to ensure that they make

the altitude restriction at CEJAY. Instead, the PF intentionally (and inappropriately) engages LVL CHG for the descent. The following performance outcomes were defined:

1. Success: PM identifies the transition to LVL CHG and switches back to VNAV or manages the restriction at CEJAY in LVL CHG.
2. Less than ideal: does not change out of LVL CHG but puts 10,000 on MCP to manage restriction at SUZSI.
3. Undesirable: NA.
4. Bad: PM fails to see LVL CHG mode and misses altitude restriction at CEJAY (altitude bust).

Challenge 2 (shortened lateral path): ATC clears the airplane direct to SUZSI: "*AirlineName* 1 cleared direct to SUZSI, comply with the speed and altitude at SUZSI." This action reduces the number of track miles of the lateral path but still requires the same descent. It can be hard to descend and slow down with the shorter route. The following performance outcomes were defined:

1. Success: PM comments on not making it or needing to descend more steeply; asking for relief.
2. Less than ideal: could have managed it better (late to VNAV or to take action).
3. Undesirable: PM comments so late that they cannot meet restriction at SUZSI.
4. Bad: Airplane fails to cross SUZSI at 10,000'; crosses high or low without ATC permission.

Challenge 3 (airspeed error): In vectoring for the approach, ATC asks the airplane to slow: "*AirlineName* 1, reduce speed to 170 knots for spacing and after PRINO intercept the 26L localizer and track it inbound." The PF intentionally fails to call for flaps 5 as the airplane slows. The following performance outcomes were defined:

1. Success: PM requests additional flaps, selects Flaps 5 prior to airspeed reaching 178 (tied to PF call).
2. Less than ideal: PM requests additional flaps but they are already below 178 and above 170.
3. Undesirable: NA.
4. Bad: airspeed reaches 170 with no PM input on need for additional flaps.

Challenge 4 (inappropriate mode): As they are cleared for the approach: "*AirlineName* 1, maintain at or above 3,800' until you are established, cleared for the ILS 26L approach," the PF intentionally fails to arm the approach (APP). The following performance outcomes were defined:

1. Success: PM ensures that arming occurs right after the clearance.
2. Less than ideal: PM prompts APP Mode prior to needing to start down on the Glideslope.
3 Undesirable: NA.
4. Bad: Flightcrew needs to use a different vertical mode to capture the G/S from above.

Appendix D. Simulator and Eye Tracking Data Visualization Application

A Data Integration and Visualization Tool for Research on Pilot Monitoring

Software Solutions Addressing Synchronization Challenges.

Understanding complex socio-technical work and its supporting human-computer interaction is an important and growing need. We address monitoring by cockpit crew flying highly automated airliners. Monitoring is the sensemaking process of understanding the dynamic flight situation. Research in this HCI domain requires tools that themselves provide good HCI design, to support data integration and visualization. Several independent data streams had been collected for the same events, which were run in an airline's training simulator. To support coding this data we identified key requirements and built a web-based tool that met our specific needs while designing for flexibility as well. This tool provides animated playback of simulator data displayed on a representation of the flight deck. It synchronizes this with eye fixation sequences projected onto the flight deck displays and with over-the-shoulder video of the pilots' activities. The tool enabled effective coding of pilots' performance monitoring challenging events.

CCS CONCEPTS • **Human-Centered Computing → HCI Design and Evaluation Methods • Visual Analytics • Activity Centered Design**

Additional Keywords and Phrases: Eye-Tracking, Pilot Monitoring, Video Analysis, Data Visualization, Web Development, Situational Awareness, Aviation Human Factors, Data Integration

1.1 Introduction

Operation of socio-technical systems that are complex, safety-critical, and highly automated poses a difficult design challenge for human-computer integration. Piloting of commercial airliners is an interesting case of such systems, and construction of the flight deck poses a challenging design problem. On the one hand, this is a well-established and intensively researched design space. On the other, the automation behind the interface has evolved dramatically and its design has become a compromise between integrating system advances while still preserving familiar legacy aspects, some elements of which are a hundred years old. Piloting critically depends on the crew gathering information from the interface to build an understanding of the rapidly changing situation. In short, pilot monitoring is foundational to effective control.

How crews make sense of the intricate array of displays on the flight deck is an important and ongoing area of research. Given increasingly complex air traffic control and increasingly automated aircraft systems, there is growing interest and concern about how well pilots monitor the flight, and how well the interface supports their activity [1]. Commercial aviation maintains an extremely high level of safety. However, inadequate monitoring has been a contributing factor to accidents, major upsets, and non-compliance with Air Traffic Control (ATC) guidance [2, 3]. Reviews of incidents in commercial aviation suggest that flight crews are sometimes unaware of deviations in even basic flight parameters, such as low airspeed [4]. Despite such slips, human control of airliners is critical for their safe operation; human pilots manage the unexpected, as happens most frequently in the dynamic phases of ascent and descent. The increasing complexity of both automation and operational environments in future airspace places further demands on human interaction with aircraft.

Thus, research on the interaction between crew and flight deck, in the process of monitoring, is an important research topic. Such research, typically carried out in simulators, benefits from use of multiple data types, and heterogenous data streams. Audio data can capture verbal communication between crewmembers and ATC. Over-the-shoulder video records gross movements of crew, whether body language or type of control action (e.g. pulling throttles back). Flight parameters and other data captured from the simulator (or aircraft) provide a detailed accounting of aircraft control inputs and resulting state, as they change over time. In addition, visual attention is central to monitoring, and eye-tracking provides information about eye-fixation and scan patterns, both important clues about attention.

From a research perspective, understanding monitoring performance is very likely to depend on manual scoring by researchers. Such scoring likely depends on multiple indicators. Monitoring is centrally a matter of attending and understanding, cognitive processes that do not necessarily have reliable, general, pre-specifiable behavioral indicators. Thus, data integration tools are vital to research. In addition, tools developed here may also assist in operational settings, such as improved support for flight instructors who are also evaluating pilot monitoring performance.

This paper concerns the development of a data integration and visualization tool to support a study on the effect of training and display configuration on pilot monitoring. This study collected a large volume of data, namely simulator log files,

over-the-shoulder video, and associated eye-tracking. The study required expert assessment to score whether the pilot appropriately monitored a sequence of challenging events. After establishing initial requirements for a synchronization tool, researchers and developers worked interactively and iteratively further improving the application.

1.2 RELATED WORK

Coordinating multiple streams of unaligned data is by no means a new problem. Multiple tools allow screen capture of a computer and may combine this with eye-tracking overlayed on the screen. The tool closest to our needs was ChronoViz, which allows the synchronization of multiple data streams particularly video and eye-tracking data. It also provides a number of helpful markup and time-stamped annotation tools [5]. A variety of eye-tracking tools allow video capture of the participant's field of view, such as the flight deck and out the window, and superimpose the eye's position sampled over time. This can be viewed in playback using a tool like ChronoViz or projected live into a headset. AugmentedEye allows an instructor to sit behind a pilot in a simulator, and through an augmented reality equipped headset, view the location of the trainee's eyes in real time [6].

There was a large gap between these existing types of systems and our needs. The video data we possessed was completely different, being over the shoulder, from the field of view captured in eye tracking. This over-the-shoulder video showed little about system state and made it difficult to discern flight parameters. Therefore, we needed to reconstruct a visual representation of the relevant parts of the flight deck. This representation provided the frame for providing the animated values of the simulator parameters. The eye-tracking trace was superimposed over this frame, post hoc. Additionally, when determining scanning patterns, where the eye 'fixates' is much more relevant than its position at every instance, something the solution described in this paper calculates before-hand, and plots accordingly.

1.3 Design and Development

Design Requirements

Most importantly, the application must integrate play-back of the simulator state, the pilot's eye fixations, over-the-shoulder video and audio. Specific requirements are listed below.

- The tool needs to create a fairly accurate animation of data describing the airplane's flight, with attention to detail such that pilots, researchers, and instructors can easily discern what is happening. To ensure accuracy, the animated display needs also to align with how the eye-tracking was captured.
- It is desirable to show values not displayed precisely on the flight deck, yet still relevant to flight, like throttle position, warning lights, or flap handle positions. This is a key advantage of animation. Video play-back does not allow explicit, easily discernable display of aircraft parameters nor display of values that are temporarily blocked, or simply excluded from the camera's field of view. Through animation, the tool can display values not on the flight deck to add to the user's understanding of airplane state.
- The tool must solve for disparities in the temporal structure of the different data streams. Even if the eye-tracking and simulator logs are collected at different rates, it must appear to the user that they are equally granular.
- Synchronization must always be preserved, even if the user hops from time to time, or chooses to step through playback frame-by-frame. The starting times of the data streams also need to be aligned.
- The tool should be quick to load, and fast on already synchronized data, as well as being lightweight in memory usage so it can be run on multiple device types.
- The user should be able to access existing event markers and add new annotations throughout play-back.
- It should be written in a language and with frameworks popular amongst programmers to be easily recognizable.
- It should never be burdensome to use, with easy control of play-back and the ability to step through pivotal moments in flight frame-by-frame.
- The tool also needed to be developed quickly, as it was critical for ongoing research, yet at the same time built with infrastructure in place to support future modifications and extensions.

Data Pre-Processing

The first task for the application is to synchronize the eye-tracking and simulator log data. Each is stored as a comma-separated values (CSV) file. The file entries correspond to the position of the eye at any given instance and variables concerning the airplane's situation respectively. Both datasets are stamped with universal time codes (UTC) from the same time server, and after correcting for a frequency mismatch, the data are merged into one table. While simulator data were collected 6 times per

second, or at a frequency of 6 Hertz, eye-tracking was recorded at 60 Hertz. To solve this, when initializing the data, the application linearly interpolates the sample simulator data to 60 hertz, duplicating each entry 10 times to maintain consistency.

Next the application needed to recognize when key events in flight occurred, like passing over a waypoint or whenever the field elevation was set on the MCP. This way, researchers could skip ahead to these critical points when playing back the simulation and observe what pilots were doing and where they were looking. A script scans through the simulator data, determining when such events took place, marking their time stamps and annotations in a separate CSV. These annotations are referred to as "Event Markers." When viewing the simulation, researchers can add annotations throughout play-back, and download them to this file.

The last step in data processing is the calculation of fixations. Instead of plotting the position of the eye at every moment, only eye fixations were drawn to better understand a pilot's scan pattern. A final script runs through the eye-tracking data, determining the velocity of each eye movement. If the eye's velocity is greater than 100 degrees per second, the script flags this as a saccade, a rapid movement of the eye, for example, between objects of interest (AoI). Otherwise, low velocity eye movements indicate a fixation, which the application will animate. This technique of separating fixations from saccades is detailed in Salvucci & Goldberg [7]. Finally, the data capture includes an indicator of gaze confidence. We used an 85% criterion, animating only fixations that have a confidence higher than this cutoff.

The Application Stack

The application is run in a web environment, and as a result is lightweight, maintainable, and easily adaptable. Written in Angular, a Typescript based framework, the code is modular with each component on the flight deck framed as a component in Angular. Components are the main building blocks for Angular applications, consisting of a hyper-text markup language (HTML) template that declares what renders on the page, a type-script class that defines its behavior and function, cascading style sheet (CSS) selector that defines the styling of the component, and its own localized data. On top of the flight deck components, there exists a 'controller' component that allows the user to play, pause, or step through individual frames of play-back. This component utilizes Angular Services to send data about the application state to other components and read rows from the initialized CSVs. With adaptability in mind, this design allows future experiments to modify existing components or add new ones with limited engineering time.

The front-end of the application uses HTML Canvas for animations, and Angular to organize components. Initializing the data will store everything the application needs on the user's disk, which is accessed by the ExpressJS backend. In this way, the aforementioned steps of pre-processing don't need to be repeated if the user closes the tool and starts it again. There is thus a sense of persistence, which can be helpful when dealing with larger datasets that take time to pre-process.

The current layout of the application was a result of iterative tweaks requested by the researchers to more accurately depict the aircraft displays. The display can additionally be scaled to any size, preserving the shape and relative sizing of each component. The eye-tracking recorded where the pilot was looking on a given component, with x/y coordinate pairs that range from (0,0), representing the bottom left, to (1,1), representing the top right. Because the data was normalized in this fashion, it makes plotting data, regardless of the size or dimensions of the screen, trivial. All that remains is the controller component updating the position of the eye or flight parameters, if and only if they change, at a frequency of 60 hertz.

Simulation and Playback

After starting the application, the user is presented with a flat display panel of the various components on the flight deck. Of key interest are the Mode Control Panel (MCP), Primary Flight Display (PFD), and Navigational Display (ND) which are animated. The FMS, EFIS and EICAS displays are not critical to this experiment, and as such are left as empty boxes. Whenever the pilot looks away from all of the given areas of interest, the 'Off AoI' box is outlined in red and the pilot's eye fixations color changes to gray. If the gaze has a confidence of less than 85%, the 'No Data' box similarly turns on. Figure 1 shows the components on the interface. Areas outlined in red display simulator values and the large area outlined in black represents any space the pilot could look.

Figure 1: Flat panel of various components. Controller component is the bottom right box.

After inputting an offset to align the video and simulator data, the user can simply press play and view all video, fixation trajectory, and simulator values in sync. Calculating the offset between the video and simulator/eye-tracking data is simple given that the UTC time is also displayed in the bottom right corner of the video.

Figure 2: Input an offset to align the video and simulator data. In this case the offset is 422 seconds.

Figure 3: Pictured above is an 'Event Indicator' which are key events in flight that pause the play back and alert the user. Other event indicators are colored bars in the controller component below. By hovering over them the user can read their annotation.

Event markers will alert the user at key points in flight. These annotations can also be added during playback.
Only the five most recent fixations are drawn, with lines drawn between them to indicate a scanning pattern. Circles are largest for the most recent fixation.

Figure 4: A close up view of a pilot's eye shifting from the Primary Flight Display to the Navigational Display.

1.4 CONCLUSION

We successfully developed a data synchronization and visualization tool with an effective human-computer interaction design. In turn, this tool supports researching issues concerning human-computer integration in the safety-critical socio-technical system of crew and flight deck, such as how system monitoring is affected by training. This tool was effectively used to code pilot monitoring performance from over 30 hours of flight in our simulator study of airline pilots.

The tool successfully integrated three heterogeneous data streams. It aligned the times of these three, while allowing the user to view and mark the synchronized event data. A key feature was providing appropriate visualization of each stream. The tool uses animation rather than video to show system state, which provides much more flexibility, precision, and control than does reliance on videos of the displays captured during the events. For example, variables captured by the sim but not shown in the flight deck can be included. Our visualization of eye-tracking shows the trajectory of fixations. In short, large heterogeneous data collections are not only of value for computer analysis, as in much machine learning, but also enable people to make complex categorizations of dynamic events, when provided with relevant, temporal visualization of the data.

The tool was developed quickly with very limited resources. Rapid development was aided by using the web environment and the Angular framework. In turn, this contributes to future adaptability of the tool. API calls to third party applications or simply adding additional data sources should be easy to engineer. We suspect it will be both straight-forward and valuable to develop layouts for other flight decks and complex interfaces, e.g., for other aircraft, for monitoring Unmanned Aerial Vehicles, or for ATC. There are several straightforward extensions. Animating the additional components for which we have data but did not animate (the Upper and lower EIAS and EFIS). would give a better understanding of why a pilot's gaze may shift there. Providing an automatic alignment of start points would remove a calibration task currently the responsibility of the user. Further, additionally parameterizing the display design should make it still easier for users who have little or no programming experience to adapt the tool to new uses. Adding subcaptioning is also an area for development. The simulator environment is very noisy, making it difficult to understand pilot conversation. Preprocessing the audio to remove noise, applying speech recognition, and displaying speech in subcaptions would be valuable. Finally, it is worth noting that while this was developed as a tool for researchers, an extended version might be helpful in operational applications, such as aiding flight instructors.

Complexity of automation and of computer-mediated work will continue to increase, in flying and in many other domains. Understanding how work unfolds over time and its current strengths and weaknesses will continue to be both challenging and important. Multiple data streams, of human physiology, of engineered system state, and of emergent activity will likely be needed to build such understanding. Continued development of tools to synchronize and visualize the data will be critical.

Acknowledgements

We would like to thank Zoey Sun for assistance in software development and Cesar Ramirez for discussions and code review. The source code for this application and additional data is available at https://github.com/JOHN-DOE/TheDashboard

REFERENCES

[1] FAA. 2013. Operational use of flight path management systems; Final report of the performance-based operations. Aviation Rulemaking Committee / Commercial Aviation Safety Team Flight Deck Automation Working Group. Washington, DC: FAA

[2] Randall J Mumaw, Dorrit O. Billman, and Michael Feary. 2020. Analysis of Monitoring Skills and a Review of Training Effectiveness. NASA Ames Research Center.

[3] Mumaw, R.J., Billman, D., and Feary, M. (2019). Factors that influenced airplane state awareness and incidents. CAST SE-210 Output 2. NASA/TM-20205010985.

[4] Airplane state awareness safety analysis team, "Airplane State Awareness Joint Safety Analysis Team Interim Report," Tech. rep., Jun. 2014.

[5] A. Fouse and J.D. Hollan. 2010. DataPrism: A tool for visualizing multimodal data. In Proc. Measuring Behavior 2010, 1:1–1:4.

[6] J.I.D. Vlasblom, J. van der Pal, and G.K. Sewnath. 2019. Making the invisible visible – increasing pilot training effectiveness by visualizing scan patterns through AR. ITEC, Stockholm. NLR.

[7] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications Symposium* (pp. 71-78). New York: ACM Press.

Appendix E. Results from Pilot Debriefing

Summary of Pilots' Final Reactions

Following the last scenario, pilots filled out a questionnaire and for 17 of 19 pilots this included verbal follow-up questions and discussion. Pilots rated five statements on a 7-point scale, where a rating of 4 was neutral and higher numbers expressed agreement. Then they answered 7 free-response questions and reviewed and discussed written responses with the experimenter.

The questions were designed to elicit critical as well as positive reactions, though there likely were demand characteristics making it easier to report positive than negative reactions. While participant ratings are not a direct measure of training effectiveness, these measures are often related, and it is valuable for the participant to experience the training as beneficial.

The free-response questions and following discussion elicited many specific, relevant assessments from pilots. These observations and suggestions, in addition to comments made during training, will be used to identify revisions to the training design used in the study. We assumed that pilots generally like simulator time. Therefore, we were particularly interested in assessing whether, and how frequently, pilots perceived there was value added beyond having additional time flying in the simulator.

Table 1 shows the rating questions and responses. Overall, pilots were quite positive about the experience, with average ratings on each question greater than 6 on the 7-point scale. No scores were on the disagree or not useful side and only one score was neutral (4). Because the ratings were so high across items, we don't have a great deal of sensitivity to see differences between items. Comparing ratings of the overall session to flying in the sim, six participants rated the whole session greater than flying the sim, while three rated the sim higher than the whole session.

Table 1. Pilot Judgements on the Rating Task: 1–7 scale

Topic	*Rating of*	*Average*	*Proportion Top-Rated*
The activities and content presented in the overall 3-hour session were	useful to me	6.58	13/17
The session will improve my monitoring	agree	6.47	10/17
The overall activities and content would be helpful as part of FO training.	agree	6.53	10/17
Flying in the simulator was	useful to me	6.47	9/17
The Principles & Practice exercises and material (in the debrief room) were	useful to me	6.21	8/17

The pilot assessment (final debrief) as well as the tutorial (mid-point debrief) had a flexible format, following up pilot questions or observations. In the final debrief (and often also in the tutorial)

several themes emerged and are mentioned here. Many additional, useful comments were provided by smaller numbers of pilots.

Broad Themes

- Pilots were very positive about the value of training that targets monitoring and/or the role of Pilot Monitoring. Many commented that this was not explicitly or extensively taught and/or said that getting the additional training on this topic was valuable.
- Pilots often commented positively about interest and usefulness of the model of monitoring. Several also wanted more practice applying it (e.g., as in the one video clip).
- Pilots frequently commented positively about the interest, realism, or value of the specific scenarios used in the sim.
- Issues of when and what to communicate from monitoring were the subject of considerable pilot-initiated discussion, both in the tutorial and in the final debrief. The tutorial included discussion about the importance and timing of communicating what you have observed, but we were struck by both the importance of and the variability in how pilots understood this. Several said they had been criticized in the PM role for (over)communicating, and they varied in whether this had been or was still an issue for them. This was widely thought to be a diminishing if not resolved issue. However, in addition, several pilots said this had been an issue when they were younger and might be an issue for new pilots, e.g., before permanent hire. Several pilots also stated they thought it should be a PF responsibility to share their plan so that the PM would know how the PF was planning on flying.

Implementation Level

- Many pilots liked interactive elements in the tutorial (such as the video clip) but wanted more of this and less text-heavy slides.
- For the sim, assessment of the confederate Pilot Flying was extremely varied, with some pilots considering the PF performance not that problematic, to a bad day, to concluding that performance this poor must be scripted.
- Several pilots thought that monitoring challenges in which the PF was scripted to make errors was very valuable for training both alert noticing and assertive communication.