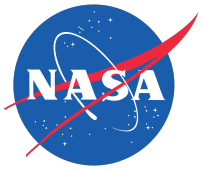
NASA/TP–20240007669



# Physiological and Subjective Responses of Pilots during Advanced Air Mobility Flight Testing with Automated Systems

*Kevin J. Monk*
*NASA Ames Research Center, Mountain View, CA*

*Tyler D. Fettrow*
*NASA Langley Research Center, Hampton, VA*

*Raquel C. Galvan-Garza*
*Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ*

*Amanda E. Kraft*
*Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ*

June 2024

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 NASA Aeronautics and Space Database and its public interface, the NASA Technical Report 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:

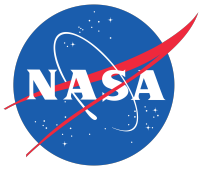
- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having 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 to *help@sti.nasa.gov*
- Fax your question to the NASA STI Information Desk at 443-757-5803
- Phone the NASA STI Information Desk at 443-757-5802
- Write to:
  STI Information Desk
  NASA Center for AeroSpace Information
  7115 Standard Drive
  Hanover, MD 21076–1320

NASA/TP– 20240007669



# Physiological and Subjective Responses of Pilots during Advanced Air Mobility Flight Testing with Automated Systems

*Kevin J. Monk*
*NASA Ames Research Center, Mountain View, CA*

*Tyler D. Fettrow*
*NASA Langley Research Center, Hampton, VA*

*Raquel C. Galvan-Garza*
*Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ*

*Amanda E. Kraft*
*Lockheed Martin Advanced Technology Laboratories, Cherry Hill, NJ*

June 2024

## Acknowledgments

This work relied on a large team. The flight test required countless hours over 2 years from Sikorsky and NASA operations, systems, and controls engineers. Lockheed Martin Advanced Technology Laboratories (ATL) supported this study by providing the functional Near Infrared Spectroscopy (fNIRS) and the Zephyr Bioharness sensors, by aiding in sensor procedure development, data collection, data analysis and interpretation. Within the Human Factors domain, we specifically acknowledge Nick Lepore and Carl Pankok (prior Sikorsky Human Factors Engineers) for helping to obtain necessary approvals and organizing the shared equipment over those 2 years. We acknowledge the contributions from Christopher S. Yang, Dan Farrell, Stacy Moran, and Benjamin McConnell for assisting with the biometric data acquisition and familiarization during this final flight test. We also acknowledge Ashima Sharma (NASA intern) for helping to learn and navigate the Tobii eye tracking analysis software and eye tracking literature during preceding familiarization flights.



Available from:

NASA Center for AeroSpace Information
7115 Standard Drive
Hanover, MD 21076-1320

National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312

Available electronically at http://www.sti.nasa.gov

# Physiological and Subjective Responses of Pilots during Advanced Air Mobility Flight Testing with Automated Systems


## Summary

Aviation is constantly evolving, mostly due to the integration of automated systems into the National Airspace System. This presents a host of challenges and opportunities. NASA's Advanced Air Mobility (AAM) project has taken a significant leap forward with a research flight test led by the Integration of Automated Systems (IAS) sub-project. This effort focused on assessing automated flight deck algorithms essential for supporting high-density Urban Air Mobility (UAM) operations. Carrying out a two-ship flight test in UAM Maturity Level 4 scenarios, the team evaluated state-of-the-art algorithms, including Hazard Perception and Avoidance (HPA) and flight path management (FPM) systems. This paper investigates into pilots' physiological and subjective responses during the flight scenarios conducted. In collaboration with Lockheed Martin Advanced Technology Laboratories (ATL), we collected and analyzed live-flight biometric data using eye tracking, mobile brain imaging, and heart rate sensors. The study provides insights into pilots' workload and cognitive engagement while navigating automated tools, leveraging both biometric data and post-encounter subjective assessments. The research highlights the interaction between human operators and automated systems, contributing valuable lessons learned about gathering human data in live-flight environments. The knowledge acquired from this study enhances our understanding of human factors in automated flight and informs future studies attempting to undertake similar feats.

# Contents



# List of Figures



# List of Tables

## Acronyms

| | |
|---|---|
| **AAM** | Advanced Air Mobility |
| **ACAS-Xr** | Airborne Collision Avoidance X for Rotorcraft |
| **AFCM** | Automated Flight and Contingency Management |
| **AOI** | area of interest |
| **ATL** | Advanced Technology Laboratories |
| **CA** | conflict avoidance |
| **CR** | conflict resolution |
| **DAA** | Detect and Avoid |
| **DWC** | DAA Well Clear |
| **EVAA** | Expandable Variable-Autonomy Architecture |
| **fNIRS** | functional near infrared spectroscopy |
| **FPM** | Flight Path Management |
| **HPA** | Hazard Perception Avoidance |
| **IAS** | Integration of Automated Systems Subproject |
| **MW** | middleware |
| **OPV** | Optionally Piloted Vehicle |
| **RA** | resolution advisory |
| **SARA** | Sikorsky Autonomy Research Aircraft |
| **UML-4** | UAM Maturity Level 4 |

# 1 Introduction

NASA initiated a National Campaign under the Advanced Air Mobility (AAM) project to address several challenges for the fusion of high-density automated aircraft into the National Airspace System. The Integration of Automated Systems (IAS) sub-project led research and development efforts related to the evaluation of flight deck automation functions needed to support AAM operations. The campaign ultimately culminated in a research flight demonstration that integrated several automated systems in a dual-ship framework that was representative of a UAM Maturity Level 4 (UML-4) environment. The primary objective of the final flight test was to test automated technologies developed under NASA's Automated Flight and Contingency Management (AFCM) sub-project: a Hazard Perception Avoidance (HPA) system and a flight path management tool Flight Path Management (FPM). Several secondary objectives included testing the improved Ground Collision Avoidance System and 4D Auto-Approach and Land algorithms, one auto-approach algorithm developed by NASA, and another auto-landing algorithm developed by Lockheed Martin Sikorsky. Another objective was to understand pilot needs regarding situational awareness maintenance and safe and timely decision-making with the research algorithms under test in live flight. This paper focuses on the characterization of the research pilots' physiological and subjective responses during the HPA and FPM scenarios from the flight test in the airspace surrounding Sikorsky's Connecticut facilities.

Lockheed Martin Advanced Technology Laboratories (ATL) became involved in this effort through ongoing collaborations with Sikorsky in developing objective pilot assessment methods and conducting human-autonomy system evaluations (e.g., on the DARPA Aircrew Labor In-Cockpit Automation (ALIAS) program, and through an Air Force Research Laboratory collaboration). ATL previously implemented subjective and objective measures of pilot workload, situational awareness, usability, and trust when working with autonomy in simulation and in flight. The functional near infrared spectroscopy (fNIRS) sensor used in this evaluation was used previously in simulated flight evaluations. However, this was the first opportunity for ATL to collect fNIRS data in flight, allowing for testing of the sensor in an operational environment. These data are important for supporting ongoing workload and other human state algorithm development efforts that are aimed at operational application.

Commonly, questionnaires and interviews are employed to gather feedback on the thoughts, workload, and attention allocation of the user. These are valuable methods, but lack the ability to collect real-time data and may also be subject to various types of reporter bias (Kruger and Dunning, 1999, Nisbett and Wilson, 1977). Biometrics provide more temporal resolution to the workload and attention allocation of the user, as well as provide a glimpse into the subconscious components of behavior (van Weelden et al., 2022). Eye tracking in particular is argued to be one of the most important biometric variables that can help identify fatigue, motion sickness, spatial disorientation, and stress or high workload (Peibl et al., 2018). Higher brain activity, as measured by fNIRS, has also been tied to increases in cognitive workload (Davies et al., 2023, Mark et al., 2024, Sun et al., 2019). However, the fusion of various sensor modalities can provide a more holistic understanding of the human's state (Causse et al., 2019, Harrivel et al., 2017, Kraft et al., 2017, Prinzel et al., 2000). Such insights will help direct the development of more broadly accessible, user-friendly, and efficient interfaces, lead to a better understanding of the more nuanced interactions between humans and automated machines, not to mention the many potential future applications of embedded real-time human state classification.

Here we present pilot biometric findings from a real-world flight test involving two live aircraft outfitted with state-of-the-art automation algorithms. Both aircraft had research pilots "in-the-loop", but here we focus on the workload and attention allocation of the research pilots flying the primary aircraft that

contained the HPA and FPM conflict resolution algorithms. There is limited research discussing biometric data collected during real flights (i.e. Di Stasi et al. (2015); Wright and McGown (2001)). The majority of the literature regarding biometrics of pilots is related to simulated flight (Feltman and Bernhardt, 2021, Gateau et al., 2015, Sun et al., 2019, Yu et al., 2020), likely for a variety of reasons. Flight tests are costly, higher risk, and difficult to achieve both technically and organizationally. Enabling the collection of biometric data during live flight is a challenge given such sensors are typically used on the ground in controlled settings. In this report, we highlight the challenges and lessons learned, as well as what collected measures can tell us about a research pilot's workload and attention allocation while aircraft perform automated procedures.

# 2 Methods

## 2.1 Test Environment

The test environment and the methods has been thoroughly described in a previous publication related to the operations (Bacchesci et al., 2024). Here we provide a brief description of the environment. The IAS group developed a framework to integrate partially to fully autonomous algorithms implemented on two Sikorsky aircraft, Sikorsky Autonomy Research Aircraft (SARA), a modified S-76B and the S-70 Optionally Piloted Vehicle (OPV), a modified UH-60. Both aircraft incorporated Sikorsky's MATRIX$^{TM}$ autonomy system, including hardware and software. The IAS team integrated several technologies through a middleware (MW) also known as Expandable Variable-Autonomy Architecture (EVAA). There was a MW instance on both SARA and OPV. The MW had socket connections between each instance to the ground, to the aircraft it resided on, and the pilot tablet onboard, which allowed for an efficient flow of data and mostly automated execution of a complex flight test (Sampson et al., 2024). SARA acted as the primary test aircraft while OPV acted as the "intruder". Figure 1 depicts the different stages of a single flight test run.

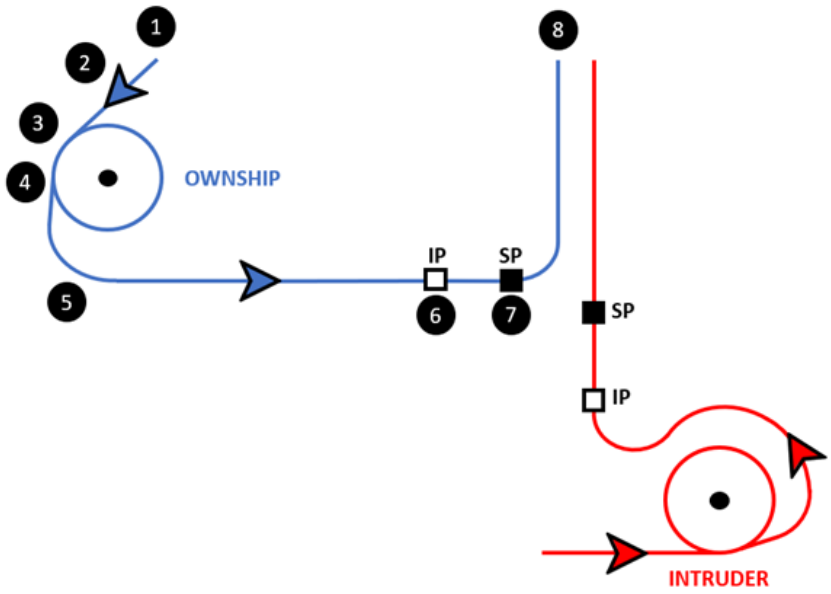



**Figure 1.** Operational Flow of flight test from Navigation to Maneuver Start Points (SP). Credit: Nelson et al. (2024)

The MW engineer on the ground, in collaboration with all researchers, decided which test card would be flown and submitted this information from a ground-based interface that modified the behavior of the pilot tablet on each aircraft. Part of this modification was providing the pilot "control" over when to initiate the subsequent flight test phase. After the MW engineer uploaded the plans (Figure 1; Step 2), the pilots would then click "Setup", initiating an autonomous flight plan towards each of their respective orbits (Figure 1; Step 3). These orbits served as the means for syncing the aircraft to allow for a precise conflict encounter in geometry and time. Once both aircraft made it to their orbits (Figure 1; Step 4), the pilot received another option on their tablet to "release" from the orbit, which started several operations that set the research starting point (T0) for each aircraft, and the research algorithm for that test card (Figure 1; Step 5). The flight to T0 was fully automated. All preceding steps are referred to as 'Setup Mode' in this report. Once the aircraft arrived at T0 (Figure 1; Step 7), the pilot tablet would automatically switch from the MW situational display to a display that was more appropriate for the algorithm under test. Depending on the algorithm under test and the test run, the pilot had to make decisions about which resolution to accept and execute, and on occasion let the algorithm make the decision and execute autonomously. This segment that involved the conflict resolutions with the systems under test are referred to as 'Research Mode' in this report.

## 2.2 Systems Under Test

As stated in the Introduction section, the primary objective of this flight test was to test and obtain data on the HPA and FPM research algorithms. The HPA algorithm was the Airborne Collision Avoidance X for Rotorcraft (ACAS-Xr), which presented alerting and guidance for Detect and Avoid (DAA) and conflict avoidance (CA) conflicts (Rorie and Smith, 2024a,b). DAA conflicts required research pilots to manually fly within the suggestive range of headings and/or altitude targets to resolve the encounter within $\leq$55 seconds of penetrating the DAA Well Clear (DWC) threshold. Once a DWC violation was no longer avoidable, CA conflicts required either pilots or the automation to immediately comply with resolution advisories (RAs) that commanded a target track for avoidance of a near mid-air collision. DAA and CA conflicts were scripted based on the intruder blundering into the ownship's flight path at varying miss distances, with offsets incorporated to preserve safety of flight. The FPM algorithm was the Autonomous Operations Planner, which provided multiple choices of conflict resolutions (CRs) in response to far-term conflicts arising in the flight path (Ballin et al., 2024a,b). In a subset of FPM cases, the research pilots simply had to monitor aircraft performance to the flight plan and scan for a potential CR. Table 1 illustrates the individual test runs included in the dataset for this report.

## 2.3 Human Factors Data Acquisition

The flight test spanned 2 weeks, and flights were conducted on weekdays that had acceptable weather conditions. There were 2 sorties per day, and each sortie (a single bout of flight) was about 120 minutes from engine start to landing. Each test run was dedicated to testing one of the two research algorithms. The amount of test runs executed within a sortie varied depending on the length of the flight or presence of technical difficulties. Before each sortie, the NASA research pilot for each aircraft would meet with NASA and Sikorsky human factors personnel to outfit the pilots with various biometric devices. Specifically, the pilots were outfitted with a mobile functional near infrared spectroscopy (fNIRS) Artinis PortaLite fNIRS (Einsteinweg, The Netherlands), Tobii Pro 3 wireless eye trackers (Danderyd Municipality, Sweden), and the Zephyr Performance Bioharness (Medtronic Zephyr, Boulder, CO, USA). The fNIRS device is a noninvasive and portable technique that uses light to measure oxygenated (HbO)

**Table 1.** Description of Flight Test Matrix.

| System Under Test | Automation Level | Res. Exec. Role | Guidance Stimuli | Pilot Response | Total Encounters |
|---|---|---|---|---|---|
| FPM | Automated | Automation (w/ consent) | Directive CR(s) | Button press | 23 |
| | No Action | N/A | N/A | Monitor Conformance | 10 |
| HPA | Automated | Automation (w/ consent) | Directive RA(s) | Button press | 1 |
| | | Automation (w/o consent) | Directive RA(s) | Monitor Automation | 12 |
| | Manual | Human | Suggestive DAA and/or Directive RA(s) | Flight Inceptors | 18 |

levels from two electrodes on the forehead. Changes in HbO provide insight on brain activity in the prefrontal cortex in real-time as oxygenated blood is redirected and consumed in areas of increased neuronal activity (Murkin and Arango, 2009). For these flight tests, fNIRS sensor was placed over the right dorsolateral prefrontal cortex. The eye tracker measures gaze position, pupil diameter, and motion of the head (i.e., accelerations and angular rates). The Zephyr Bioharness measures heart rate and breathing measures, along with accelerometry and posture. Together these biometric devices provide a multi-faceted observation of the pilot's continuous and dynamic physiology.

The SARA pilot was presented with a post-encounter questionnaire in-flight at the conclusion of each test card. The questionnaire automatically popped up when the pilot clicked "Stop" on the MW display on the tablet. Questions probed pilots' subjective ratings of workload and ride quality during each scenario, and pilots responded by using their finger to drag the sliders to the most appropriate rating for the encounter that was just completed. The workload rating was based on a modified Bedford Workload scale (Roscoe, 1984) (1 - Insignificant to 10 - Impossible), and ride quality was rated on a scale of 1 (Very Smooth) to 10 (Very Rough) (Figure 2). Acceptability of CR presented by the algorithms was also probed within the same window, and responses to these queries are discussed in separate publications dedicated to evaluations of each system under test (Ballin et al., 2024a, Rorie and Smith, 2024b). The questionnaire was designed to take less than one minute.

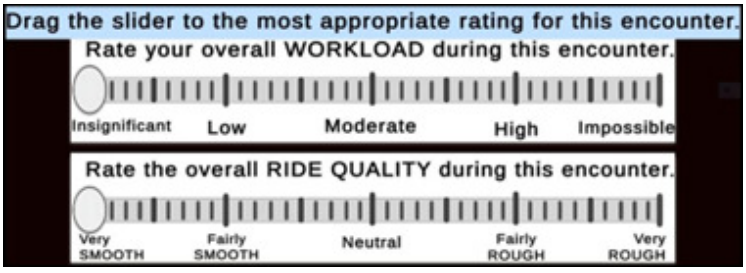



**Figure 2.** In-flight Post-Encounter Questionnaire - Pilot Workload and Ride Quality Assessments.

## 2.4 Data Processing

The raw eye tracking files were initially processed through the Tobii Pro Lab software to generate automated mappings of gaze to specific areas of interest (AOIs) over time. AOIs included the research tablet (Figure 3), flight instruments, test cards, and the out-the-window view. For each AOI, we computed the duration of time spent looking within the AOI as well as the ratio of time looking within the AOI to all other mapped gaze. Mean gaze step was computed as the average distance (in pixels) traveled between consecutive gaze points. Fixations and saccades were classified by the Tobii Pro software, from which we derived duration of fixation/saccade, number of fixations exceeding 300ms, and the mean saccade speed. Pupil diameter, rotational head velocity, and linear acceleration of the head were also provided by the eye tracker sensors.

**Figure 3.** Research Tablet AOI.

Artinis OxySoft software was used to compute oxygenated hemoglobin (HbO) concentrations using the Modified Beer-Lambert Law, using the first 10s of each recording as the hemoglobin concentrations baseline. The exported data was then manually reviewed to determine if either channel or significant periods of the recording needed to be excluded due to excessive noise. Artifact rejection was applied at the recording-level and segment-level based on interquartile range (IQR), as adapted from the Homer2 MATLAB toolbox for fNIRS data processing. The biometric data was further cleaned using a low pass filter and outlier removal for data points exceeding 2.7 standard deviations from the mean (Jahani et al., 2018). For each of the two channels, we computed the mean, median, standard deviation, minimum, maximum, and sum, as well as the slope over a given time window and the change ($\triangle$) in channel magnitude between the beginning and end of the window (Causse et al., 2019). The fNIRS measures are differences from baseline, such that positive values indicate increased activity or higher effort, while negative values indicate lower levels of exertion than during the baseline period.

Zephyr's summary export file includes a variety of physiology and activity measures, along with device status and system confidence estimates. Heart rate, breathing rate, and heart rate variability (HRV) were pre-processed to exclude data flagged system error values and zeroes. Heart rate and HRV were further restricted to cases where heart rate confidence was greater than 50 (out of 100). For each measure, we computed the same standard measures as described for fNIRS.

## 2.5 Statistical Analysis

We derived many outcome variables from each device's data stream; however, we do not report all of the variables here. We chose a single variable to represent heart rate and brain activity data, due to significant correlation among measures within a modality. The measures derived for eye tracking covered distinct data types (e.g., pupil, movement, AOIs), therefore we included multiple dependent variables for eye tracking. All non-eye tracking variables were quantified within the research mode segment of interest. A variable that has a $\Delta$ refers to a change within the time segment from beginning to end of that segment, and a positive value indicates the measure is higher at the end of the segment. All other dependent variables are averaged over the time segment. The primary time segment of interest in the results is Research Mode unless stated otherwise.

**Table 2.** Measurements.

| Data Type | Measurement |
|---|---|
| Workload | 1:Insignificant - 10:Impossible |
| Ride Quality | 1:Very Smooth - 10:Very Rough |
| Heart and Respiration | Heart Rate $\Delta$<br>Breathing Rate $\Delta$<br>Heart Rate Variability $\Delta$ |
| Brain Activity (fNIRS) | HbO $\Delta$ |
| Eye Tracking | Pupil Diameter<br>Saccade Frequency<br>Saccade Duration<br>Attention Allocation on AOIs<br>Dwell Time (fixation duration)<br>Gyrometry |
| Aircraft Dynamics | Aircraft Pitch<br>Aircraft Roll<br>Aircraft Groundspeed<br>Aircraft Airspeed |

**Table 3.** Experimental Manipulations.

| Experimental Manipulations | Levels |
|---|---|
| Automation Level | No Action vs. Auto vs Manual |
| Display Mode | Setup vs. Research |

Statistical analyses included separate general linear models for each measure to test differences across each level of the experimental manipulation (see Table 2 for measurements and Table 3 for experimental manipulations). The Display Mode refers to the time segments of interest, specifically Setup Mode (Figure 1; steps 1-7) and Research Mode (Figure 1; steps 7-8). The Display Mode variable is only relevant to the eye tracking data, as the other measures are only reported for the Research Mode segment that required explicit pilot engagement. The Automation Level was treated as a categorical variable. We calculated the range (maximum-minimum) for each variable relating to aircraft dynamics within each time segment. Additionally, univariate and correlation analyses were conducted to characterize the relationship between all measures during research scenarios. All statistical analyses were

performed in IBM SPSS Statistics version 29.0.2.0. Significant results at an alpha level of 0.05 are reported where appropriate. The low sample size of eye tracking data during manual resolutions did not allow for pairwise comparisons between varied levels of automation. Descriptive statistics are reported for subjective workload and ride quality ratings.

## 3 Results

Figure 4 displays the sample sizes for the devices of interest. The sample size refers to the number of test runs that were completed with the respective device. Heart rate contained the most successful test runs at 64, followed by the in-flight questionnaires with 57. The fNIRS had 39 successful test runs, while eye tracking had the lowest sample size with 23 successful test runs. Unsuccessful completion of a measurement was caused either because the device wasn't worn or there were technical difficulties. The following sections provide results for each of the respective collection devices.

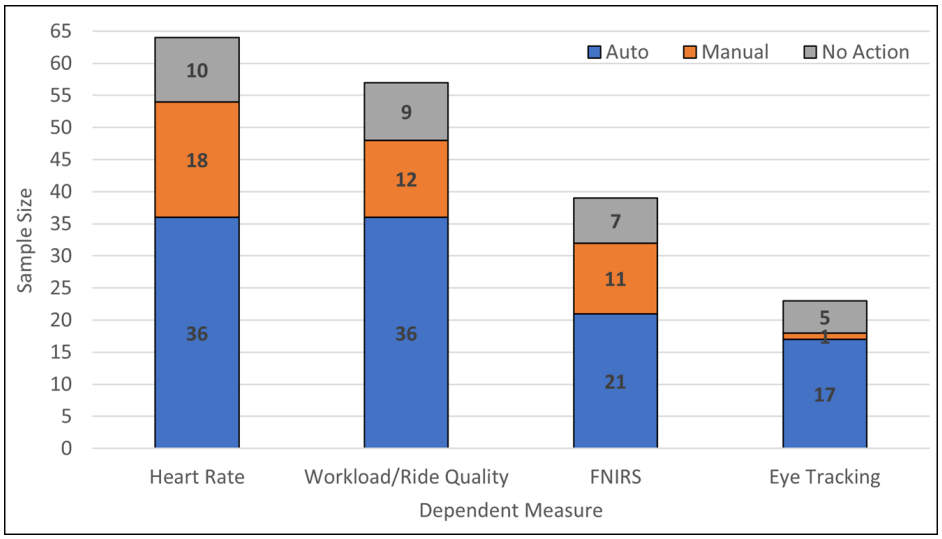
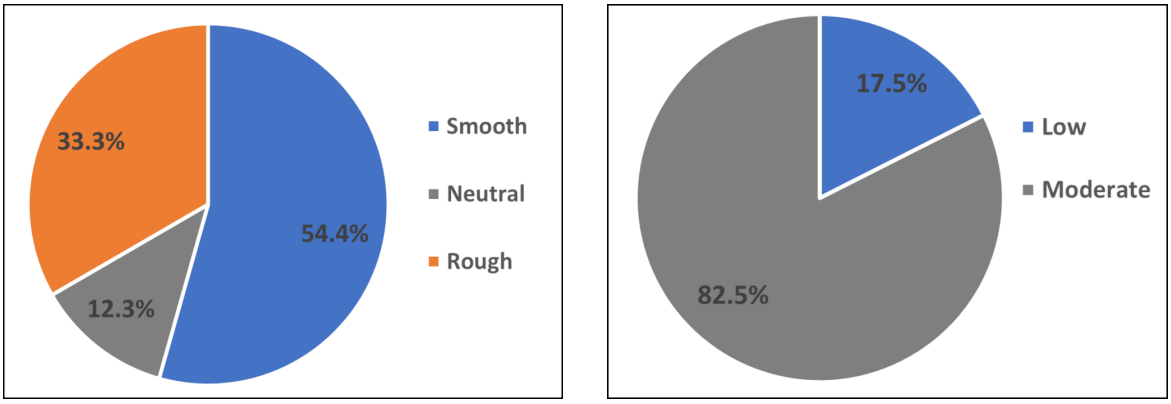



**Figure 4.** Measurement Sample Sizes by Automation Level (N = 64).




**(a)** Workload Ratings. **(b)** Ride Quality Ratings.

**Figure 5.** Workload and Ride Quality Ratings Results.

### 3.1 In-flight Questionnaires: Subjective Workload and Ride Quality

Overall, 10 of 57 encounters (18%) were rated as 'Low' workload ($\leq$3 on the modified Bedford scale) and 47 of 57 encounters (82%) were rated as 'Moderate' (4$\geq$7 on the modified Bedford scale; Figure 5ba). The range of responses for workload was between 3 and 7 ($M$ = 4.5). Ride quality assessments revealed 31 of 57 encounters (54%) were rated as 'Smooth' ($\leq$4), 19 encounters (33%) were rated as 'Rough' ($\geq$6), and 7 encounters (12%) were rated as 'Neutral' (5; Figure 5bb). The range of responses for ride quality was between 2 and 8 ($M$ = 4.5). Only four of the encounters received a roughness rating of 7 or above, with the roughest rating of 8 being given to an encounter that contained multi-dimensional CRs. Ride quality was rated as slightly smoother for encounters with 'Low' workload ($M$ = 3.56) compared to 'Moderate' workload ($M$ = 4.8). However, the relationship between workload and ride quality was not statistically significant ($r$ = .18, $p$ > .05).

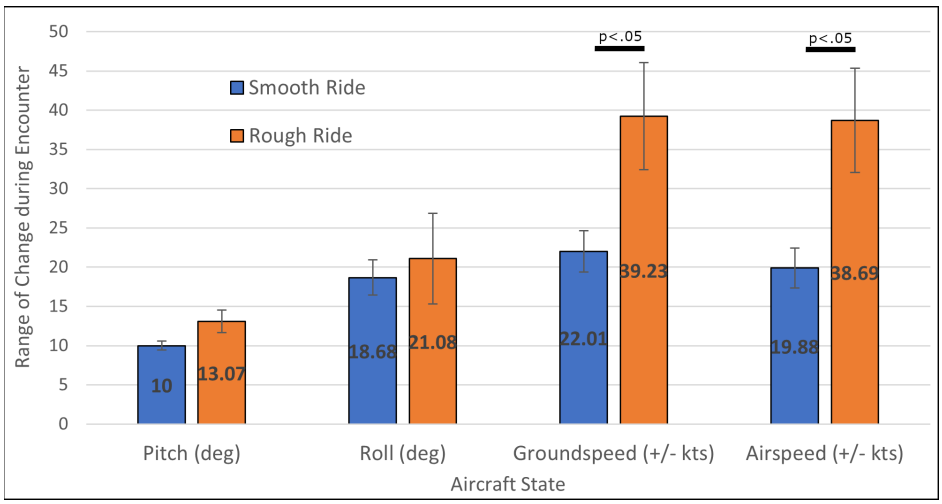



**Figure 6.** Average Aircraft State Changes by Subjective Ride Quality.

Automation Level did not impact subjective workload ratings; there was only a 0.54 mean difference in ratings between manual and automated encounters, and a 0.09 mean difference between automated encounters and encounters without any action taken, $p$ > .05. While none of the aircraft dynamics variables significantly correlated with subjective workload ratings ($p$'s > .05), aircraft airspeed range did correlate positively with raw ride quality scores, $r$ = .272, $p$ < .05. When assessing the nature of test encounters that led to non-neutral ride quality ratings, groundspeed ranges were significantly greater during encounters that were rated as 'Rough' ($M$ = 39.23kts, $SE$ = 6.82kts) compared to 'Smooth' ($M$ = 22.01kts, $SE$ = 2.62kts), $F$(2,55) = 3.62, $p$ < .05. Airspeed ranges were also significantly greater during encounters that were rated as 'Rough' ($M$ = 38.69kts, $SE$ = 6.63kts) compared to 'Smooth' ($M$ = 19.88kts, $SE$ = 2.55kts), $F$(2,55) = 4.07, $p$ < .05 (Figure 6). With regard to Automation Level, automated resolutions ($M$ = 5.18, $SE$ = 0.30) received higher ratings on the roughness scale relative to manual resolutions ($M$ = 4.01, $SE$ = 0.49) on average. Specifically, resolutions executed manually by the human pilot via flight controls ($M$ = 4.01, $SE$ = 0.49) received smoother ride quality ratings than resolutions that were automatically executed by the automation without pilot permission/human input ($M$ = 5.77, $SE$ = 0.53), $p$ < .05 (Figure 7). The correlation between ride quality and subjective workload ratings was non-significant, $r$ = .18, $p$ > .05.

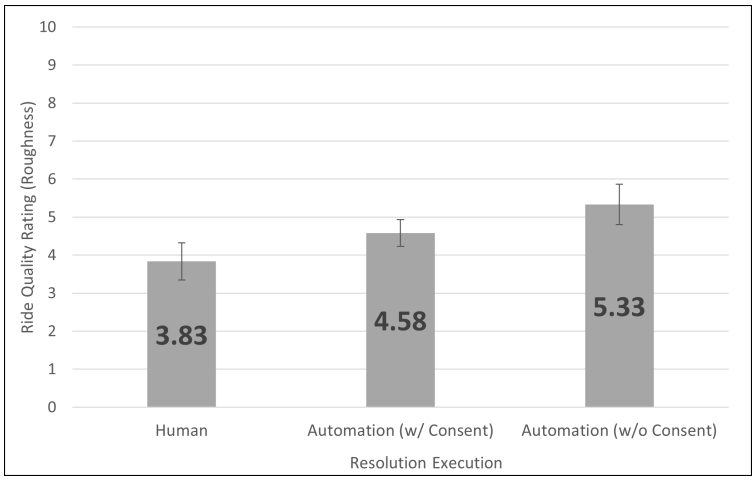



**Figure 7.** Ride Quality Ratings by Resolution Execution.

## 3.2 Brain Activity

No main effect of automation level or display mode found on fNIRS data. Brain activity positively correlated with the pilot's total number of fixations ($r$ = .53) and amount of time spent dwelling on test cards ($r$ = .70) during a given scenario, $p$'s < .05. Conversely, brain activity decreased as their mean pupil diameter ($r$ = -.55) and focus on flight instruments increased ($r$ = -.56), $p$'s <.05 (Figure 8). There was no positive correlation found between brain activity and percent of time spent focusing on any areas of interest (including the research tablet). It should be noted that any relationships between fNIRS and eye tracking data only applied to 25% or less of the test encounters overall due to limited samples. Brain activity was positively correlated with increased changes in groundspeed ($r$ = .35) and airspeed ($r$ = .34) during conflict resolutions, but the correlations between the variables were non-significant after accounting for a single extreme outlier in the speed data that was more than 3 standard deviations above the 75th percentile, $p$'s > .05. A non-significant negative correlation trend was observed between fNIRS data and subjective workload ratings, $r$ = -.27, $p$ > .05.

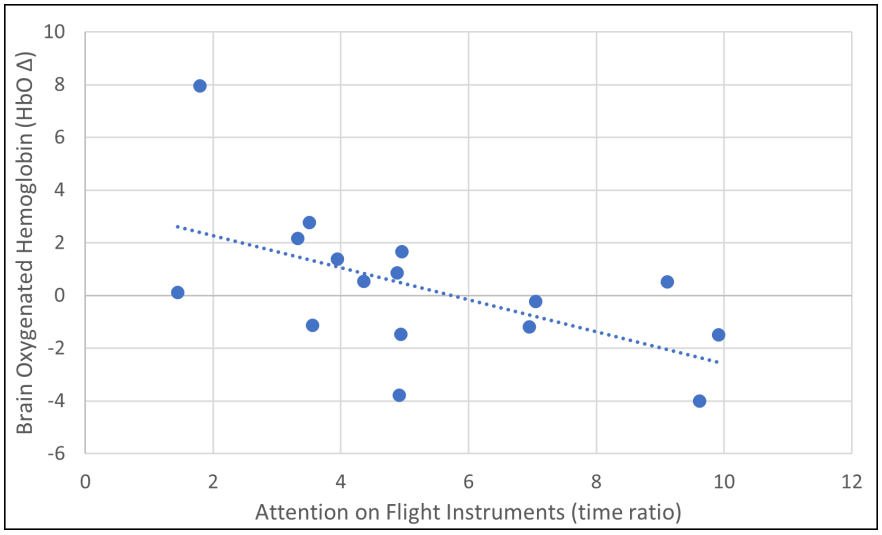



**Figure 8.** Pre-frontal Cortex Activity by Attention on Flight Instruments ($n$ = 16).

### 3.3 Heart Rate

Neither breathing rate nor heart rate variability were significantly impacted by any of the experimental manipulations, $p > .05$. No main effect of automation level or display mode was found on heart rate data, $p$'s > .05. Heart rates were also elevated by 7bpm during scenarios where ride quality was subjectively rated by pilots as 'Rough' ($M$ = 5.19bpm, $SE$ = 2.41bpm) compared to 'Smooth' ($M$ = -1.47bpm, $SE$ = 0.93bpm), $F(2,58) = 3.87$, $p < .05$ (Figure 9). Heart rates during encounters rated as 'Neutral' (5) did not significantly differ from the $\Delta$s observed during either Rough or Smooth rides. The correlation between heart rate and subjective workload was negligible, $r = .09$, $p > .05$. Increased ranges of pitch changes by the aircraft tended to elevate heart rate ($r = .30$), but the correlation between the variables was non-significant after accounting for a single extreme outlier in the pitch range data that was more than 3 standard deviations above the 75th percentile ($p > .05$).

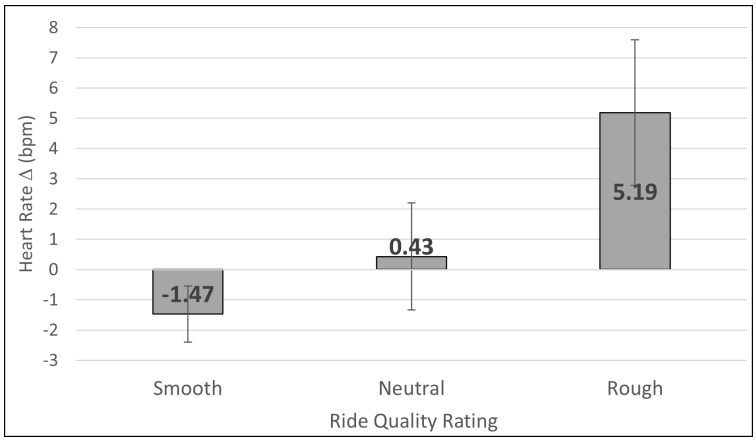



**Figure 9.** Heart Rate $\Delta$ (Pre-to-Post Encounter) by Ride Quality Rating.

### 3.4 Eye Tracking

There was a main effect of Display Mode on multiple eye tracking metrics. The research pilot allocated a higher percentage of visual attention to the flight instruments during Setup mode ($M$ = 8.08, $SE$ = .825) compared to Research Mode ($M$ = 5.22, $SE$ =.825), $F(1,46) = 6.00$, $p < .05$. However, the pilot spent a higher percentage of time focusing on the research tablet during Research mode ($M$ = 9.91, $SE$ = 0.99) compared to Setup mode ($M$ = 5.40, $SE$ = 0.99), $F(1,46) = 10.19$, $p < .05$. Saccade frequency was greater in Research mode ($M$ = 7.44, $SE$ = 0.29) compared to Setup mode ($M$ = 6.25, $SE$ = 0.29), $F(1,44) = 8.10$, $p < .05$. Average pupil diameter was also greater in Research mode ($M$ = 2.32, $SE$ = .03) compared to Setup mode ($M$ = 2.18, $SE$ = .03), $F(1,46) = 8.79$, $p < .05$. With regard to head motion, the pilot looked up/down (gyroX) more frequently during Research encounters ($p > .05$) and side-to-side (gyroY) significantly more often during the Setup ($p < .05$). While in Research mode, a significant inverse relationship was found between mean saccade duration and research tablet focus, $r$ = -.43, $p < .05$. As indicated in Figure 10, greater degrees of roll changes during conflict resolutions also tended to increase the amount of attention allocated outside the window of the aircraft, $r = .49$, $p < .05$. None of the variables had a statistical impact on dwell time. Samples did not allow for a main effect of automation level on attention allocation, but the one singular case that included eye tracking data for a Manual scenario (executed by pilot via inceptors) yielded significantly less visual attention to the research tablet compared to the Automation and No Action scenarios.

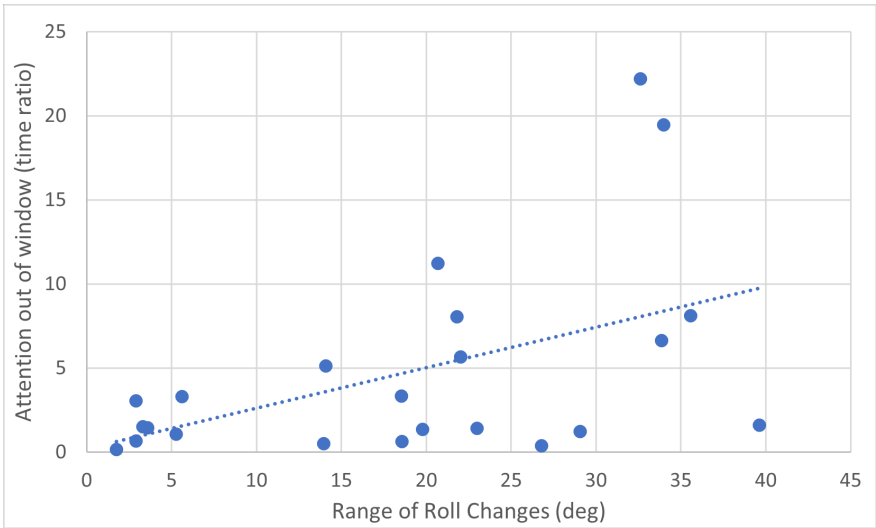



**Figure 10.** Attention Out of the Window by Change in Aircraft Roll.

# 4 Discussion

All test runs were safely and successfully executed to the researchers' satisfaction. The research pilot on SARA consistently reported manageable workloads throughout testing. The biometric device setup was straightforward, and the questionnaire software performed reliably during flights. We did encounter some difficulties with the biometrics devices: participants occasionally chose not to wear the eye-trackers and the fNIRS device, and there were some reports of discomfort. Nevertheless, the study offered valuable lessons learned and the data sheds light on pilot workload and attention allocation while flying automated aircraft algorithms, culminating in a successful human factor evaluation.

## 4.1 Subjective and Physiological Implications

The study revealed a lack of correlation between subjective workload and ride quality, indicating they were influenced by distinct factors. It is essential to note that automated systems executed most resolutions throughout the study and pilots never reported 'High' workload in any flight test run. The most time- and safety-critical encounters were rated as 'Moderate' workload, at worst. According to the revised Bedford scale, pilot workload was never rated low enough to be considered 'Insignificant' (rating of 1), but also never progressed to a level high enough to be considered intolerable for the task. On average, pilots faced a manageable workload. Thus, even the procedural tasks that were done manually did not create substantial burden on the pilots' perceived workload. Although pilots were pre-briefed on what to expect throughout all encounters for safety reasons, the consistently low-to-moderate subjective workload is a credit to the research systems under test since ease of use is an inherent and intended benefit of the resolution decision aids and automated execution performed by the algorithms. This is especially encouraging when you consider the ambiguous real-time transparency of the logic behind the resolutions and a subset of encounters included rare, worst-case scenarios that required immediate avoidance of an imminent near midair collision.

These subjective measures also showed a weak relationship with objective physiological measures such as brain activity and heart rate. The biometric data correlated more closely with specific aspects of aircraft performance than subjective ratings. Interestingly, significant changes in aircraft speed were aligned with brain activity, and aircraft pitch changes corresponded with heart rate. Overall, there was

no statistically significant correlation between the biometric data and subjective perceptions of workload or ride quality. However, a main effect between heart rate and subjective ride quality was observed when the range of responses were binned by category, as heart rates increased during encounters rated as a 'Rough' ride after they were clear-of-conflict. Subjective workload was basically identical regardless of the amount of pilot input required, but automated resolutions without the pilot in-the-loop resulted in subjectively rougher ride quality - presumably due to the increased speed changes and unpredictable nature of a management-by-exception style of execution by the automation during Resolution Advisories. Other potential factors include the pilot acceptability of the user interfaces, research tablet, magnitude of automated resolutions, and pitch oscillations that were induced when automation was engaged; all of which are extensively explored in several separate publications stemming from this flight test (Bacchesci et al., 2024, Ballin et al., 2024a, Eggum et al., 2024a,b, Rorie and Smith, 2024b).

Pilot attention usually remained within the cockpit when wings were level, but as roll increased the pilot was more likely to look out of the window during encounters. This was likely to maintain spatial orientation during aircraft heading adjustments. Pilot attention was increasingly directed toward flight instruments during Setup mode where the primary responsibility was to monitor the automation's continuous changes to the aircraft states when setting up the upcoming research encounter. Decreased brain activity while looking at instruments may indicate that the pilots were in a monitoring role as the automation achieved the commanded aircraft states before and after the decision-making component of the conflict assessment.

The transition to Research mode increased focus onto the research tablet where traffic information and CR algorithms were located. The higher frequency of eye movements between areas of fixation in Research mode indicate that there were more information sources being simultaneously monitored during research scenarios. Saccades were more frequent in Research mode, and the duration of these saccades decreased while attention was focused on the research tablet. This correlation suggests more erratic attention switching between individual display elements on the tablet itself. These findings are reasonable considering that even in cases where automation executed the resolution, subsequent pilot action or input was potentially required based on continuously updating alerting and guidance on the algorithm displays on the research tablet.

One pilot noted that the absence of aural alerting during the flight path management scenarios (which accounted for 78% of eye tracking cases) caused them to shift attention to the research tablet in anticipation of an imminent time-critical resolution change more frequently and longer than they normally would during crosschecks. This extra focus was evident in the data as pitch motions of the head ('gyroX') were greater in Research mode due to the heads-down display being out of their central focal area; however, a secondary analysis did not reveal a statistically significant difference in research tablet attention between the two systems under test. Ideally, a heads-up cockpit display would have aided the crosschecking of multiple information sources, but the current test environment did not appear to degrade task performance. Greater yaw motions of the head ('gyroY') in Setup mode is indicative of more attention shifting to the safety pilot in the left seat as lip reading may have aided effective communication in the noisy cockpit. Lastly, although average pupil diameter has been linked to mental effort in addition to light exposure in previous human-computer interaction literature (Mathôt, 2018), it negatively correlated with brain activity and raw workload scores in the current study. The authors caution that several limitations (e.g., small sample size) should be considered when inferring any repeatability of these results, and these considerations are further discussed in the following section.

## 4.2 Limitations and Lessons Learned

There were numerous challenges associated with the in-flight data gathering, such as ensuring safety, device portability, battery longevity, user comfort, minimizing distraction, and securing requisite approvals. Here we discuss these challenges and limitations, followed by suggestions for improving these limitations in future studies.

In our study, we addressed the safety concern with a dedicated "safety pilot" who maintained ultimate control of the aircraft, allowing the "research pilot" to focus on the algorithms and manual control when necessary. We mitigated battery life issues with an onboard AC port that provided endless powering capabilities, but this was only feasible for the eye-tracking device. However, multiple factors limited the sample size of eye tracking data, in contrast with the more consistent acquisition of brain activity and heart rate data. Firstly, the motion within the cockpit along with the bulky flight suit occasionally caused an inadvertent disconnection of the power supply cord during sorties that extended beyond the battery's life capacity. There was also an instance where the recording unit shut down during a flight where the pilot attempted to place it in their pocket for improved stability, which may have led to overheating. The pilot comfort and potential for distraction posed persistent hurdles. One research pilot initially expressed severe discomfort with the eye trackers while wearing the required pilot helmet (pressure behind the ears), which prompted the researchers to modify the arms of the glasses to alleviate the issue. The other research pilot opted out of using the eye tracker altogether due to incompatibility with their prescription eyeglasses.

The data quality, in particular the eye tracking data, was lower quality relative to data collected during pre-flight simulation runs. Factors such as helicopter vibrations, variable lighting conditions, and the pilot's adjustments of the devices for comfort contributed to the sub-optimal data. The impact of lighting on the eye tracker's accuracy and pupil size further complicates the reliability of the data. Real-world testing often limits the extent to which these challenges can be improved. Thus, future research using Tobii Pro Glasses would also benefit from the acquisition of the Protective Lens Add-on designed to mitigate the negative impacts of very bright or dark environments on eye-tracking data quality.

The fNIRS device faced its own set of in-flight challenges, including shifts along the forehead that compromised data quality and posed a potential distraction to the pilot. The fNIRS device's battery life did not always survive through the full duration of a sortie, as the devices began collecting data before pilots began walking to aircraft for the step and pre-flight checklists. The fNIRS device is also confined to monitoring brain activity in the prefrontal cortex, which presents notable limitations. Moreover, individual variations in brain anatomy mean that the same electrode placement might sample disparate brain areas across different individuals (Taubert et al., 2020). Given that we're examining a singular brain region in just a couple of subjects, individual distinctions in cerebral structure and activity likely bear greater significance. Further, age and experience are associated with different brain activity patterns (Iordan et al., 2020, Reuter-Lorenz and Cappell, 2008) - factors that are critical when interpreting modest outcomes of the brain data in our research.

The approach to data analysis in this study could have obscured significant results. By averaging out or selecting temporal data points for each dependent variable and using minimum and maximum values for aircraft variables within designated time frames, we potentially filtered out the impact of brief, high-stress incidents, such as automation induced oscillations or CRs. These events may have only lasted 15 seconds within a Research mode extending over two minutes. Given the sensitivity of biometrics to even subtle task demands, our current approach on analysis may have diminished the temporal precision of the biometric data, thus lowering the likelihood of detecting these transient, yet

impactful stressors.

The primary objective of the flight test was to evaluate research automation algorithms, with the human factors components we are discussing here being considered ancillary. Thus, the flight test's success was not contingent on the human factor biometric data. When a device failed to collect data for the reasons discussed above, we did not repeat test cards, therefore we forfeited the biometric data associated with that test point. Despite these challenges, we have acquired multiple lessons learned from this experience:

- In-flight Questionnaire Design: The vibratory conditions of the aircraft cockpit was sub-optimal for completing precise touch-based questionnaire tasks on an unmounted and unfavorably sized tablet (Bacchesci et al., 2024). Future flight test research that emphasizes the probing component could benefit from employing force feedback technology (i.e., gravity wells) in unstable environments, as it has improved aimed movements and target selection on next-generation aviation displays in previous research (Monk et al., 2015).
- Advanced Workload Measurements: Future research studies with more time flexibility for questionnaires would benefit from a multi-dimensional workload scale such as the NASA Task Load Index to distinguish the weight of physical vs. cognitive demand within pilots' subjective workload scores when assessing correlations with ride quality.
- Expand Attention Areas of Interest: Our AOIs were limited and did not encompass the entirety of the cockpit, leading to underrepresented data for certain interactions. In the future, capturing or generating an image of the full cockpit would allow classifying gaze as in-cockpit and taking the inverse as out of window for higher accuracy. This approach would also allow generation of a global coordinate system for mapping gaze rather than gaze relative to the wearable eye tracker. Furthermore, there are different components within the instrument panel and the research display tablet that could be further divided to understand exactly what kind of information the pilot is acquiring within each of the AOIs.
- More controlled environment: Keep things like test cards and the tablet as stationary as possible with a secure mount (without constricting performance/comfort). It is difficult to maintain a real-world environment and yet establish an environment that is suitable for all methods of biometric data acquisition. However, for attention tracking, stationary objects are easier to identify. This may be easier for things like a tablet that only require a 2D interface, but for something like a card deck with multiple cards to sift through, making this stationary is very difficult and at times difficult to track what the participant was tracking.
- Measuring Light and Use of Protective and Corrective Lenses: Tobii's protective lenses would have benefited eye-tracking data quality in the variable-lighting test environment. As mentioned earlier in this section, lighting not only effected the quality of the data (ability to track the eyes), but it likely also confounded pupil size when the data was acquired. Protective lenses may help in these situations. Moreover, acquiring lighting data (how much light is getting to the pupils), would allow for statistically controlling for the lighting conditions. Additionally, one pilot chose not to wear the eye trackers due to the need of corrective lenses. Providing embedded corrective lenses may have substantially increased the sample size for this data type.
- Balance Pilot Participation: Generate more structured pilot assignments to test runs to data evenly distributed across types of test runs as best as possible. In this study, only one pilot wore the eye

trackers, and coincidentally did not perform any manual conflict resolution test runs. This limited the types of analysis we could do with the eye tracking data.

- Continuous Data Analysis: Future analyses should try to analyze in a more continuous manner as opposed to averaging over large time frames. This can enrich both retrospective interpretation and offer potential benefits for real-time analysis, aiding researchers and users alike.

# 5 Conclusion

The current study presents unique aeronautic research, as it intertwines pilot biometric analysis with live automated flight operations. Under NASA's Advanced Air Mobility (AAM) project, the Integration of Automated Systems sub-project has led the charge in assessing necessary automation features for seamless integration into the National Airspace System. Completing a two-ship research flight indicative of an Urban Air Mobility Maturity Level 4 environment, this campaign tested pivotal technologies like the Hazard Perception Avoidance (HPA) system and flight path management (FPM) tool. By studying pilots' physiological and psychological responses to these advanced systems, we have garnered valuable insights regarding situational awareness and decision-making capacities. Lockheed Martin ATL's collaboration introduced a new dimension to the study, employing state-of-the-art sensors in an operational environment. Recording pilots' gaze, brain activity, and heart rate responses, in conjunction with subjective feedback, offered a more holistic narrative of the human-machine interactions. The present study not only highlights the lessons learned, but also showcases our commitment to pioneering human research in automated systems. The human is still a part of an automated system, and it's important to refine methods of data collection and analysis to aid the development of future interfaces between machines and their operators.

## References


Bacchesci, N., Eng, D., Nwajagu, E., Saravanakumaar, R., Scofield, J., Williams, E., Fettrow, T., and Monk, K. (2024). Integration of automated systems test campaign. AAM-NC-129-001.

Ballin, M., Barrows, B., Nelson, S., Karr, D., Fettrow, T., and Wing, D. (2024a). Flight evaluation of in-flight strategic path planning automation for future high-density operations. NASA Technical Publication - In Porgress.

Ballin, M. G., Barrows, B. A., Nelson, S. L., Underwood, M. C., Fettrow, T., and Wing, D. J. (2024b). Flight evaluation of a flight path management system for high density advanced air mobility. AIAA Conference 2024.

Causse, M., Chua, Z. K., and Rémy, F. (2019). Influences of age, mental workload, and flight experience on cognitive performance and prefrontal activity in private pilots: a fnirs study. Scientific Reports, 9.

Davies, H. J., Williams, I., Hammour, G., Yarici, M., Stacey, M. J., Seemungal, B. M., and Mandic, D. P. (2023). In-ear spo for classification of cognitive workload. IEEE Transactions on Cognitive and Developmental Systems, 15:950–958.

Di Stasi, L. L., Diaz-Piedra, C., Suárez, J., McCamy, M. B., Martinez-Conde, S., Roca-Dorda, J., and Catena, A. (2015). Task complexity modulates pilot electroencephalographic activity during real flights. Psychophysiology, 52:951–956.

Eggum, S., Zahn, D., Williams, E., Patterson, G., and Monk, K. (2024a). Flight test evaluation of autonomous descending-decelerating precision point-in-space aproach to the ground. In Progress.

Eggum, S., Zahn, D., Williams, E., Patterson, G., Saravanakumaar, G., and Ringleberg, R. (2024b). Test evaluation of automation-induced oscillations. In Progress.

Feltman, K. A. and Bernhardt, K. A. (2021). Measuring the domain specificity of workload using eeg: Auditory and visual domains in rotary-wing simulated flight.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., and Dehais, F. (2015). Real-time state estimation in a flight simulator using fnirs. PLoS ONE, 10.

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., Abraham, N. A., Prinzel, L. J., Motter, M. A., and Pope, A. T. (2017). Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing.

Iordan, A. D., Cooke, K. A., Moored, K. D., Katz, B., Buschkuehl, M., Jaeggi, S. M., Polk, T. A., Peltier, S. J., Jonides, J., and Reuter-Lorenz, P. A. (2020). Neural correlates of working memory training: Evidence for plasticity in older adults. NeuroImage, 217:116887.

Jahani, S., Setarehdan, S. K., Boas, D. A., and Yücel, M. A. (2018). Motion artifact detection and correction in functional near-infrared spectroscopy: a new hybrid method based on spline interpolation method and savitzky–golay filtering. Neurophotonics, 5:1.

Kraft, A., Russo, J., Krein, M., Russel, B., Casebeer, W., and Ziegler, M. (2017). A systematic approach to developing near real-time performance predictions based on physiological measures.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77:1121–1134.

Mark, J. A., Curtin, A., Kraft, A. E., Ziegler, M. D., and Ayaz, H. (2024). Mental workload assessment by monitoring brain, heart, and eye with six biomedical modalities during six cognitive tasks. Frontiers in Neuroergonomics, 5.

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function.

Monk, K. J., Strybel, T. Z., Vu, K. P. L., Marayong, P., and Battiste, V. (2015). Effects of force feedback and distractor location on a cdti target selection task. Procedia Manufacturing, 3:2395–2402.

Murkin, J. M. and Arango, M. (2009). Near-infrared spectroscopy as an index of brain and tissue oxygenation. British Journal of Anaesthesia, 103.

Nelson, S. L., Ballin, M. G., Barrows, B. A., Underwood, M. C., Wing, D. J., Sturdy, J. L., and Williams, E. R. (2024). Designing a flight test of a flight path management system for advanced air mobility research. American Institute of Aeronautics and Astronautics (AIAA).

Nisbett, R. E. and Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. Journal of Personality and Social Psychology, 35:250–256.

Peibl, S., Wickens, C. D., and Baruah, R. (2018). Eye-tracking measures in aviation: A selective literature review.

Prinzel, L. J., G, F. F., Scerbo, M. W., Mikullka, P. J., and Pope, A. T. (2000). A closed-loop system for examining psychophysiological measures for adaptive task allocation. The International Journal of Aviation Psychology, 10:393–410.

Reuter-Lorenz, P. A. and Cappell, K. A. (2008). Neurocognitive aging and the compensation hypothesis. Current Directions in Psychological Science, 17:177–182.

Rorie, R. C. and Smith, C. L. (2024a). Flight test evaluation of the airborne collision avoidance system x for rotorcraft. DASC 2024.

Rorie, R. C. and Smith, C. L. (2024b). Hazard perception and avoidance results from the advanced air mobility project's integration of automated systems flight test. NASA Technical Memorandum - In Progress.

Roscoe, A. (1984). Assessing pilot workload in flight. Royal Aircraft Establishment Bedford.

Sampson, P., Williams, E., and Scofield, J. (2024). Synchronizing test encounter entry for multiple autonomous aircraft. In Progress.

Sun, J., Cheng, S., Ma, J., Xiong, K., Su, M., and Hu, W. (2019). Assessment of the static upright balance index and brain blood oxygen levels as parameters to evaluate pilot workload. PLoS ONE, 14.

Taubert, M., Roggenhofer, E., Melie-Garcia, L., Muller, S., Lehmann, N., Preisig, M., Vollenweider, P., Marques-Vidal, P., Lutti, A., Kherif, F., and Draganski, B. (2020). Converging patterns of aging-associated brain volume loss and tissue microstructure differences. Neurobiol. Aging, 88:108–118.

van Weelden, E., Alimardani, M., Wiltshire, T. J., and Louwerse, M. M. (2022). Aviation and neurophysiology: A systematic review.

Wright, N. and McGown, A. (2001). Vigilance on the civil flight deck: incidence of sleepiness and sleep during long-haul flights and associated changes in physiological parameters. Ergonomics, 44:82–106.

Yu, B.-W., Jeong, J.-H., Lee, D.-H., and Lee, S.-W. (2020). Detection of pilot's drowsiness based on multimodal convolutional bidirectional lstm network. pages 530–543. Springer International Publishing.