

SUBSCALE FLIGHT TRAINING EFFECTS ON PILOT WORKLOAD AND PERFORMANCE DURING VTOL HQTES

Antonio Chappell[1], Sachithra Karunathilake[2], Anthony Comer[1] & Pratima Saravanan[2]

[1]School of Mechanical and Aerospace Engineering, Oklahoma State University, Tulsa, OK, USA
[2]School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK, USA

Abstract

This paper examines incorporating flight testing as a complement to simulator-only training for remotely piloted fly-by-wire vertical takeoff and landing (VTOL) aircraft. Ten participants were assigned to two groups and completed three Froude-scaled vertical-flight Handling Qualities Task Elements (HQTEs): vertical reposition, pirouette, and hovering turn-and-hold. Both groups first trained using a validated digital twin of the S2F-X200 lift-plus-cruise VTOL aircraft in the X-Plane 12 environment using a fixed first-person-view (FPV) camera matched to the flight-test configuration. Group A completed two simulator training sessions, while Group B completed one simulator training session followed by one matched subscale flight-test session. Both groups subsequently completed a common final evaluation in simulation. Objective performance metrics, Cooper-Harper ratings, NASA-TLX workload scores, and eye-tracking measurements were collected to assess performance, workload, and visual attention across sessions. Results show that, relative to simulation, subscale flight testing was associated with higher task demands, including increased off-axis stick activity and degraded intermediate-session task performance. However, final simulator performance remained broadly comparable between groups, and participants exposed to flight testing generally perceived greater training improvement. These results suggest that subscale flight exposure may provide training value not captured by simulator-only practice by exposing pilots to real disturbances, consequence, and compensation demands while preserving a controlled simulator-based evaluation environment.

Keywords: VTOL, Handling Qualities, Flight Testing, Flight Simulation, Human-Machine Interaction

1. Introduction

Simulation-based training is a cornerstone of modern pilot instruction because it enables safe, repeatable practice with controlled task difficulty and immediate feedback. However, the relationship between simulator “fidelity” and training effectiveness is not monotonic: training transfer depends on which task-relevant perceptual, cognitive, and motor processes are engaged, and increased physical fidelity does not automatically yield improved learning outcomes [1]. For fly-by-wire aircraft, this issue is particularly salient because the pilot primarily interacts with the closed-loop response characteristics of the flight control system (FCS). If the FCS response type and pilot command mapping are held consistent, then differences in training outcomes may be driven less by nominal vehicle scale alone and more by how the training environment shapes workload, cue utilization, and compensation strategy.

A persistent limitation of simulator-only training is the difficulty of reproducing the operational salience of consequence. Real flight can elicit different psychophysiological responses than simulated flight for comparable mission profiles, suggesting that stress, arousal, and perceived consequence can meaningfully modulate pilot behavior and learning [2]. Experimental evidence from aviation human factors further indicates that structured exposure to stressors can improve pilot performance during stressful flying tasks [3, 4]. In parallel, eye-tracking and related workload measures have been used in aviation

https://doi.org/10.71945/icas2026_0685

studies to quantify attention allocation, visual scanning behavior, and workload during dynamic flight tasks [5, 6]. These findings motivate the hypothesis that introducing a higher-consequence training exposure—while maintaining acceptable safety and cost—may complement conventional simulator practice, especially for precision vertical flight tasking.

Subscale flight testing offers an underexplored intermediate training environment that preserves real-world disturbances and consequences at substantially reduced cost and risk relative to full-scale flight. Subscale aircraft have long been used as research platforms for flight dynamics and control development, but their use as a deliberate pilot training intervention remains comparatively limited in the open literature [7, 8]. For a training study, the objective is not to infer full-scale aerodynamics from subscale data, but to provide a physically embodied environment in which the pilot flies a response-type FCS under real disturbances, task constraints, and time pressure. To ensure task comparability, scaled handling qualities task elements can be constructed using principled scaling approaches and validated task definitions. Recent work applying Froude-based scaling to ADS-33-style mission task elements (MTEs) for small VTOL UAS provides a practical precedent for scaling course geometry and timing while acknowledging that not all similarity parameters can be matched simultaneously [9]. Accordingly, this paper presents a controlled human-subject experiment that uses a common simulator evaluation environment to isolate the effect of adding flight-test exposure as a training intervention. The effect of scale is not the core focus of this paper as both the simulation and flight-test experiments were performed at the same scale. Participants completed a battery of scaled vertical flight handling qualities task elements (HQTEs) derived from proposed powered-lift VTOL HQTE definitions [10]. All participants completed simulator training and a final simulator evaluation using the same task set, standardized FPV-style viewpoint, and consistent control/interface mapping. One group completed simulator-only training, while the second group completed matched training that included flight-test exposure on a physically realized scaled course at the Skyway 36 Drone Port (hereafter “Skyway 36”) in Tulsa, OK, USA. The study objective is to quantify differences in performance, control strategy, subjective workload, and visual attention associated with subscale exposure, thereby assessing whether subscale flight testing can complement simulator-only training for fly-by-wire vertical flight tasking.

2. Testbed and Control Framework

The study was performed using a common subscale VTOL testbed and a corresponding simulation digital twin to support piloted evaluation under consistent pilot command mappings and vehicle response characteristics. This section summarizes the vehicle platform, FCS architecture, and pilot interface at a level sufficient to support interpretation of the training study; additional implementation details are provided in prior publications by Comer et al. [11, 12].

2.1 Vehicle Platform and Digital Twin



Figure 1 – The S2F-X200 lift-plus-cruise VTOL vehicle (left) & its X-Plane 12 digital twin (right)

The S2F-X200 lift-plus-cruise (LPC) VTOL configuration (Fig. 1), designed and built at the Simulation to Flight Applied Research Laboratory at Oklahoma State University (OSU), served as the test vehicle for the subscale flight-training sessions. The vehicle is a “4+1” quadplane configuration, featuring

four lifting propulsors and a single cruise propulsor mounted aft of the tail. A validated simulation-to-flight workflow was used to construct a hardware-consistent digital twin of the vehicle and to enable piloted evaluation in X-Plane 12 [11]. In the present study, the simulator implementation served as the controlled final evaluation environment, while the physical subscale vehicle provided a higher-consequence flight-test exposure intended to complement simulator practice.

2.2 Flight Control System Overview

The study is motivated by the premise that, for fly-by-wire operation, the pilot primarily interacts with the closed-loop response characteristics of the FCS rather than directly with the open-loop vehicle dynamics. Accordingly, the simulator and flight-test implementations used the same Explicit Model Following (EMF) + Trajectory Control System (TCS) control architecture and pilot command mappings employed in prior subscale VTOL flight testing [11, 12]. This architecture was developed around simplified vehicle operations (SVO), with the intent of reducing pilot workload and training time for novice or minimally trained operators. At a high level, normalized pilot inputs are converted into axis-level commands and tracked using an integrated FCS composed of trajectory control, EMF inner-loop stabilization, inverse propulsor models, control allocation, and effector-to-hardware mappings [12]. Because all HQTEs in this study were performed in vertical flight mode (VFM), transition-dependent command mapping changes were not required. As the design of the FCS architecture is not the direct focus of this paper, details regarding specific control loops are omitted for brevity.

2.2.1 Pilot Inceptors and Command Mappings

Participants used a consistent two-stick setup consisting of: (i) a spring-centered left-hand stick (heave, yaw) and (ii) a spring-centered right-hand stick (surge, sway), each shown in Fig. 2. The inceptors generate four normalized pilot input channels, $(\delta_{ver}, \delta_{dir}, \delta_{acc}, \delta_{lat}) \in [-1, 1]$, which are mapped to the commands from the SVO/TCS framework demonstrated in [12].

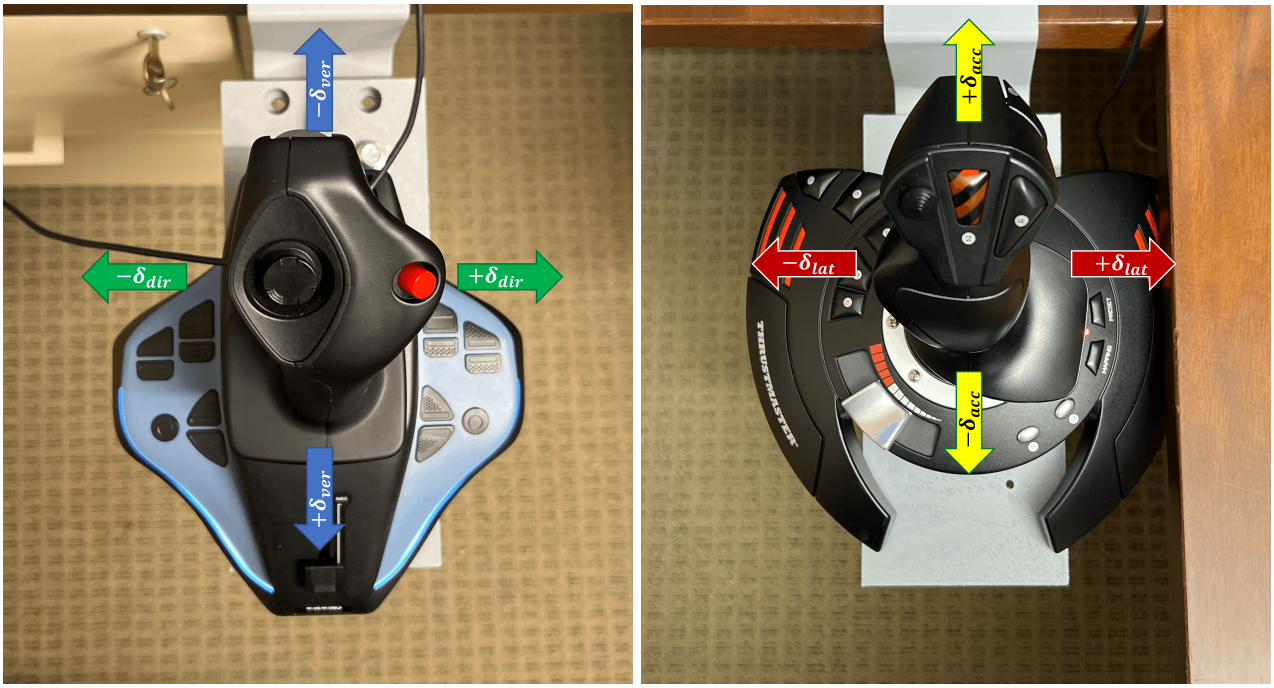


Figure 2 – Inceptor layout (left, right) in the simulation room with axes labeled

In VFM, the vertical channel commands climb/descent rate VV_{cmd}, the lateral channel commands lateral velocity $V_{LAT,cmd}$, the directional channel commands yaw rate r_{cmd}, and the acceleration channel commands normalized fore/aft acceleration a_{cmd} with a velocity-dissipation effect (described further in [12]). To reduce noise sensitivity and overcontrol while maintaining consistent control feel, all axes use a first-order input filter with time constant τ_δ and a 5% dead zone, consistent with prior subscale implementations [12, 13]. The filter time constant was set to $\tau_\delta = 0.15$ in simulation and $\tau_\delta = 0.05$

in subscale flight testing to account for control-link and video-streaming latency in the physical test environment. Inceptor command mappings are summarized in Table 1.

Table 1 – Pilot inceptors and command mappings used in both simulation and flight tests

Physical Inceptor Axis	Pilot Input	VFM Command	Maximum Command
Left stick fore/aft	δ_{ver}	Vertical velocity VV_{cmd}	200 fpm
Left stick left/right	δ_{dir}	Yaw rate r_{cmd}	30 deg/s
Right stick fore/aft	δ_{acc}	Fore/aft accel a_{cmd}	0.2 g
Right stick left/right	δ_{lat}	Lateral velocity $V_{LAT,cmd}$	5 kts

2.2.2 Automatic Holds

The FCS architecture features a set of automatic holds to aid the pilot when the inceptors are centered (no pilot inputs). The holds include: position hold, altitude hold, heading/track hold, and speed hold. These holds are described further in [13], with a summary provided in Table 2.

Table 2 – Overview of the different holds and control strategies for each respective axis

Control Axis	Hold Type	Flight State Criteria	Control Type	Command
Lateral (δ_{lat})	Position hold	$V < 3$ kts	PID	ϕ_{cmd}
Vertical (δ_{ver})	–	$h < 3$ ft	–	–
	Altitude h hold	$h > 3$ ft	PI	VV_{cmd}
Directional (δ_{dir})	Heading ψ hold	$V < 10$ kts	PI	r_{cmd}
Acceleration (δ_{acc})	Position hold	$V < 3$ kts	PID	a_{cmd}

3. Piloted Study Design

This work evaluates how the addition of flight-test exposure influences subsequent simulator handling-qualities performance and human response relative to simulator-only training. Ten participants were recruited and randomly assigned to one of two groups (A or B; $n = 5$ per group), as summarized in Table 3. Both groups completed the final evaluation in the same simulation environment (X-Plane 12) to avoid introducing additional scale-, weather-, or vehicle-dependent confounds at test time. Participants were adults aged 18 years or older, and Institutional Review Board (IRB) approval was obtained prior to recruitment and data collection (IRB-26-218).

Table 3 – Experimental design for Group A and Group B participants

Group	Training Session 1	Training Session 2	Final Test Session
A	Simulator	Simulator	Simulator
B	Simulator	Subscale flight test	Simulator

3.1 Session Structure (Training and Test)

To minimize variability due to inconsistent familiarization while avoiding unintended training advantages, all participants received standardized pre-test training materials consisting of a short instructional video and an annotated inceptor mapping document. The materials described the pilot inceptors, control-channel functions, pilot viewpoint, task objectives, and basic vehicle operation using consistent phrasing and without discussion of experimental hypotheses or comparisons across simulation and flight testing. This approach follows established piloted evaluation practice in which standardized instructional content is used to control participant exposure prior to data collection [14]. Each session (training or test) followed the same high-level structure to control for acclimation effects:

1. **Training materials (pre-brief):** Participants review a short video and a written guide describing the inceptor functions, control axes, pilot view, subjective questionnaire (Appendix C) and task objectives (described further in Sec. 3.3).

2. **Familiarization ("free flight"):** 10 minutes of unstructured flight time in the session environment, consistent with prior piloted HQ evaluation practice [14].
3. **Control comprehension ("sanity check"):** Short single-axis commands (e.g., climb/descend, left/right translation, fore/aft translation, yaw left/right) to verify that participants understand the mapping and can safely proceed.
4. **HQTE repetitions (performed in a randomized order):**
   - *Training sessions:* no fewer than 5 and no more than 10 practice runs of randomized tasks.
   - *Test session:* no fewer than 5 and no more than 10 runs of randomized tasks.
5. **Post-run ratings:** After each HQTE run, participants complete Cooper-Harper and NASA-TLX workload ratings [15], along with additional short feedback regarding perceived training utility.

3.2 HQTE Task Set

The task battery consisted of three Froude-scaled vertical-flight HQTEs spanning altitude capture, direction capture, and coupled translation–yaw control behaviors. Task intent and descriptions are adapted from proposed powered-lift VTOL HQTE definitions [10]. The selected tasks and primary measures were: (i) vertical reposition & hold (altitude error; heading error), (ii) hovering turn & hold (heading error at capture; position drift), and (iii) pirouette (ground track error; heading error). The scaled HQTE course geometry and timing are derived using a Froude-based kinematic scaling approach with a characteristic length scale. Let the geometric scale factor be $\lambda = L_{\text{sub}}/L_{\text{ref}}$, implying $L_{\text{sub}} = \lambda L_{\text{ref}},\ V_{\text{sub}} = \sqrt{\lambda}\, V_{\text{ref}},\ t_{\text{sub}} = \sqrt{\lambda}\, t_{\text{ref}}$. This scaling is used to preserve relative task aggressiveness for the subscale dynamics while acknowledging that not all similarity parameters can be matched simultaneously. The scaled course definitions were implemented in both (i) the simulator (X-Plane 12) and (ii) a physically realized scaled course at Skyway 36 for the subscale training intervention with one change: $t_{\text{sub}} = t_{\text{ref}}$. This change is made to preserve the maneuver time experienced by pilots training in subscale for the purposes of subjecting pilots to equivalent time periods of stress as would be seen at the full scale. The hypothesis behind this decision is that the effectiveness of subscale training depends, in part, on equating time under stress across scales.

3.3 Course Setup

The Froude scaling is performed based on the full-scale vehicle design (approximately 600 lbs) discussed further in [16]. The S2F-X200 is a geometric 1/3 scale of the full-scale S2F-X600, resulting in a Froude scale value of $\lambda = 1/3$. Using this scaling factor, the objectives for adequate and desirable performance on each task element are outlined in Table 4, Table 5, and Table 6. The metrics from which the original values were derived and subsequently scaled are detailed in [17], with one notable change to the pirouette course. A second reference circle was added to provide the pilots (using a fixed FPV view) with sufficient visual cueing to adequately discern their longitudinal (fore/aft) displacement.

Vertical Reposition	**Desired**	**Adequate**
Maintain longitudinal and lateral position within ±X from the hover point.	± 3 ft	± 6 ft
Maintain start/finish altitude within ±X.	± 3 ft	± 6 ft
Maintain heading within ±X.	± 5 deg	± 10 deg
Complete maneuver within X.	≤ 25 sec	≤ 30 sec
PIO tendencies in the capture and hold.	No undesirable motions that impact task performance	No divergent PIO (out-of-phase oscillations)

Table 4 – Vertical reposition (VTOL) performance requirements chart

It should be noted that, while pilot performance is only evaluated across the measures outlined in the previous section, subjects were nonetheless presented with reasonable targets for their performance.

This was intended to prevent skew in the performance evaluations in the case that a pilot would focus primarily or only on the evaluated measures.

Pirouette	**Desired**	**Adequate**
Maintain a selected reference point on the aircraft within ±X ft of the circumference of the circle	±3.3 ft	±5 ft
Maintain altitude within ±X ft:	±3.3 ft	±5 ft
Maintain heading so that the nose of the aircraft points at the center of the circle within ±X deg:	±10 deg	±15 deg
Complete the path within:	$\leq$ 60 sec	$\leq$ 80 sec
Achieve a stabilized hover (within desired hover reference point) within X seconds after returning to the starting point.	5 sec	10 sec
Maintain the stabilized hover for an additional X sec:	5 sec	5 sec

Table 5 – Pirouette performance requirements chart

Turn-and-Hold	**Desired**	**Adequate**
Maintain longitudinal and lateral position within ±X from the hover point.	±3 ft	±6 ft
Maintain altitude within ±X	±3 ft	±6 ft
Stabilize the final rotorcraft heading at the 90-degree point and 270-degree point within ±X.	±5 deg	±10 deg
Complete maneuver within X.	$\leq$ 50 sec	$\leq$ 60 sec
PIO tendencies in the capture and hold.	No undesirable motions that impact task performance	No divergent PIO (out-of-phase oscillations)

Table 6 – Turn-and-hold performance requirements chart

A digital twin of the HQTE task set was created in X-Plane 12 using a custom airport created in World Editor (WED) at the Skyway 36 Drone Port (referred to as Downtown Airpark airport [18] in X-Plane 12). Geographic data from the locations of the courses placed at Skyway 36 were used to place course digital twins in simulation.
The Froude-scaled pirouette course was created in SolidWorks and exported to Blender as shown in Fig. 3. An extension was used to generate the assets for simulation in X-Plane 12 [19]. The final simulation and flight-test realizations of the pirouette course are shown in Fig. 4.

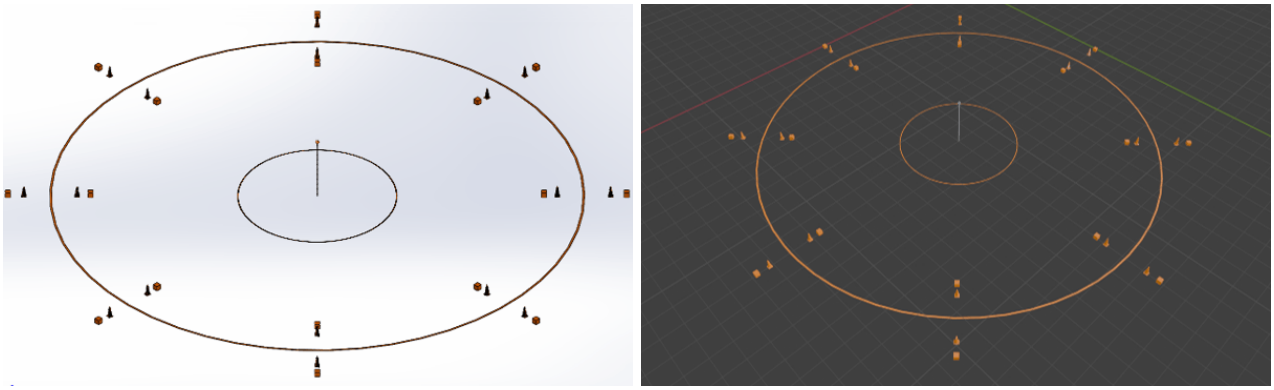


Figure 3 – Subscale pirouette digital twin in SolidWorks (left) & Blender (right)

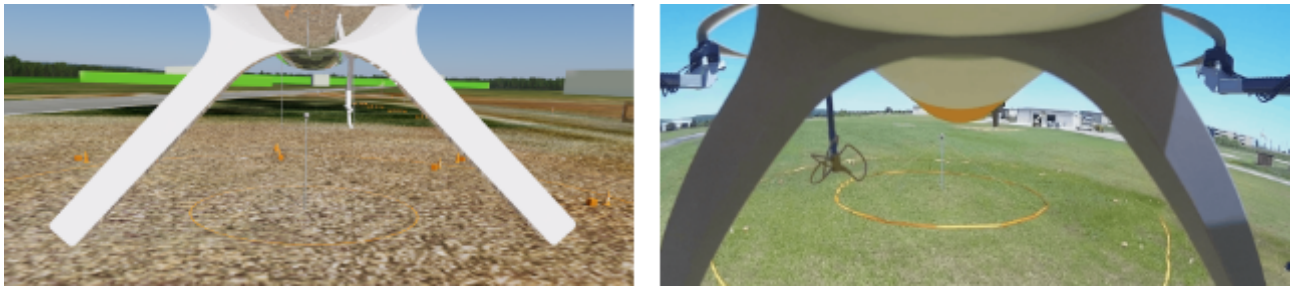


Figure 4 – Subscale pirouette in the X-Plane (left) and at Skyway 36 (right)

The same process was repeated for the creation of a digital twin for the vertical reposition and turn-and-hold courses, with final realizations in simulation and flight-test shown in Fig. 5. Following the suggestions from Chakraborty et al. [20], these courses were combined to save modeling and construction effort, and to increase uniformity and familiarity across each of the HQTEs. All Froude-scaled courses were positioned at the same locations in Skyway 36 and simulation by GPS measurement. One notable difference between the simulation and flight-test environment was elevation variation, which was not identical between setups. The X-Plane 12 digital twin utilized a flat airport as a simplification. It is possible that these differences may have slightly altered the relationship between the reference objects, instrument readings, and performance evaluations. Elevation at Skyway 36 varied $\pm$ 1 ft across each course location. In order to further align instrument readings across both simulation and subscale flight test, the orientation of each of the courses was matched. The vertical reposition, the takeoff location for pirouette, and the starting position for the left-hand, counter-clockwise turn-and-hold were all oriented due North.



Figure 5 – Subscale vertical reposition in X-Plane (left) and Skyway 36 (right)

The development of digital twins for the subscale geometry followed the same methodology outlined by Comer et al. [21]. Further details regarding the simulation model development have been omitted for brevity.

3.4 Pilot Setup

A dual-monitor setup was employed as shown in Fig. 6. The left monitor contains a pilot display, showing sensor information on aircraft states, battery voltage, flight time, and information regarding the status of the task element and pilot-in-command. The right monitor showed the FPV-style display (as depicted in Fig. 4 and Fig. 5). A Walksnail/CADDX Avatar HD system was used to provide the FPV video feed to the pilot in flight testing, while a “weapon camera” placed at the mounting location in X-Plane 12 rendered a representative display in simulation. The Walksnail system averaged a 30 ms latency throughout the flight testing.

3.4.1 Pilot Display Design

The pilot display provides accurate information about the vehicle state that may not be reasonably or precisely known through the FPV-style video feed. To ensure intuitive feedback, the display provides feedback that aligns with the pilots’ axes of control. Lateral and longitudinal velocity correspond directly with the pilots’ inputs in δ_{lat} and δ_{acc} respectively. Precise heading adjustment using the

heading indicator, given a wide horizontal span, allows pilots to pinpoint their exact heading both in degrees and at each cardinal direction. An altimeter provides similar feedback for pilot input in δ_{ver}. Additionally, similar information such as FCS hold indicators, pilot in command, battery state, and vehicle state (arm/ kill) are grouped together to provide an intuitive visual interface. These features, along with the mission status indicator, allow for the pilot and safety pilot to communicate effectively during both simulation and flight-test to ensure safe and proper operation of the vehicle.

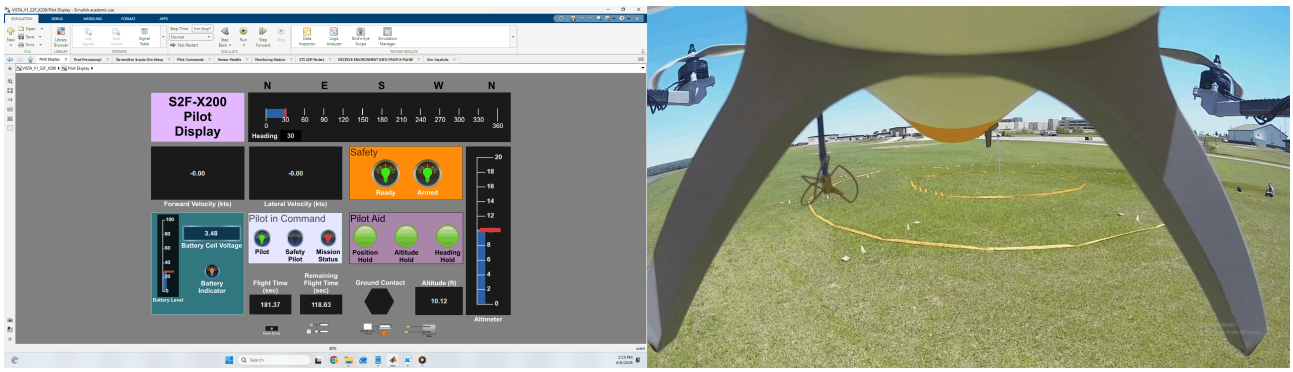


Figure 6 – Screen capture of the flight test dual-monitor setup

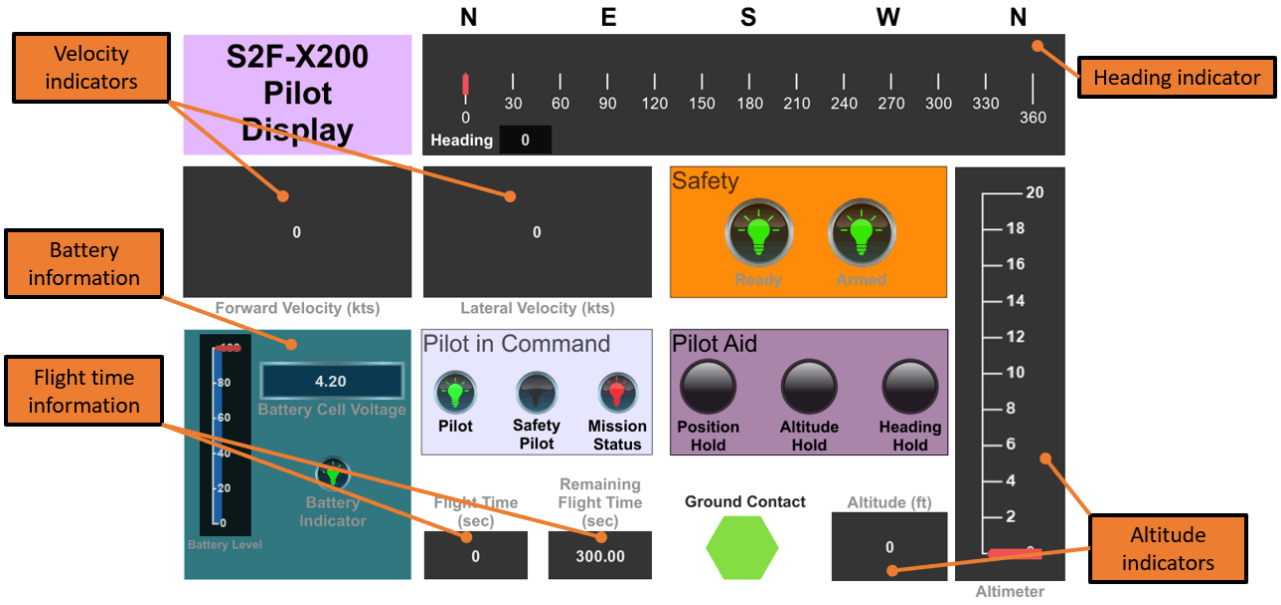


Figure 7 – Pilot display diagram

3.5 Measurements and Data Collection

To capture both performance outcomes and human response, the study combines objective vehicle metrics, eye tracking, and subjective ratings. Prior work highlights the value of wearable-based physiological sensing for stress/workload inference and the importance of aligning subjective workload instruments with the task context [22, 23, 24, 25]. Eye-tracking has also been used to relate attention allocation to pilot workload and task demand in rotorcraft-relevant contexts [26, 27], with gaze-entropy and related metrics providing a compact descriptor of visual scanning behavior as task complexity changes [28].

Upon arrival, participants provided informed consent and completed a demographic questionnaire (including prior flight, simulator, RC/FPV, and relevant gaming experience). Participants then wore the Tobii Pro Glasses 3 [29]; the eye tracker was calibrated and a brief baseline physiological and gaze measurement were collected. Eye-tracking metrics were collected at a sampling rate of 100 Hz and extracted from the Tobii Pro Labs software. Initially, consecutive invalid pupil diameter samples

from both eyes were identified and interpolated using a linear interpolation approach. The mean pupil diameter for each trial and session was obtained from the raw pupil diameter values of each eye. Pupil diameter can be responsive to luminance, task-relevant cognitive load, and arousal [30, 31]. Therefore, it is essential to obtain a baseline measurement before presenting a stimulus to the participant [32]. This study obtained the participants' baseline pupil diameter for 5 minutes prior to the start of the training session. All pupil diameter plots were generated using the baseline-corrected pupil diameter measurements as shown in Appendix A. The heat maps were generated for each trial based on the absolute duration of gaze fixation. The Tobii-Glasses are worn continuously across training and test sessions, with short breaks between sessions to mitigate fatigue.
During each HQTE run, simulator and flight-test telemetry were recorded at a sampling rate of 50 Hz and 15 Hz respectively. Telemetry data includes sensor feedback from the onboard Pixhawk Cube Orange+ flight controller, pilot inceptor inputs, and maneuver start and end signals. Participants provide post-run ratings (Cooper-Harper [33, 34] and NASA-TLX [15]). The full protocol is designed to be completed in approximately two hours, and participants are compensated for their time.

Table 7 – Primary measurement categories collected during training and test sessions.

Category	Instrument / Source	Examples of Variables
Performance	Simulator/subscale telemetry	Position tracking error, boundary violations, control activity metrics
Eye tracking	Wearable eye tracker [29]	Fixation duration, saccade rate, pupil diameter, gaze dispersion/entropy [26, 27, 28]
Subjective	Post-run surveys	Cooper-Harper [34], NASA-TLX [15], perceived training utility

4. Results

To give a sense of performance progression for each Group (A and B) within the study, representative plots of the best performing candidates in each corresponding HQTE course are described subsequently. Red lines indicate the first training session, Yellow lines indicate the second training session, and green lines indicate the final "test" session. To assess how pilot performance changed throughout each session, the root-mean-square (RMS) error outside of desired boundaries for each task is plotted. This means that all desired performances were evaluated the same, at zero RMS error. This provides a quantitative measure of the overall performance of a specific run. These "learning curves" allow comparisons to be made between individual pilots and each of the two pilot groups (A and B). The average stick activity in the second session (simulation vs. flight-test), taken to be the RMS value in each axis, is assessed to highlight differences across simulation and flight-test for all pilots. All results should be considered with the understanding that the atmospheric conditions present throughout the vast majority of the flight-testing sessions, notably a persistent and high-speed (5-10 kts) North-Westerly wind, differed from the low-speed (3-5 kts) Northerly wind conditions modeled in simulation. Additionally, some pilots highlighted differences in the ergonomics of the left inceptor across simulation and flight test, potentially exacerbating coupling between δ_{ver} and δ_{dir} in flight testing. Future work will aim to improve the consistency of atmospheric conditions and ergonomics between simulation and flight-test. Because the flight-test telemetry data collected was computed from raw GPS latitude and longitude, the resolution of the flight-test position data was on the order of 1 ft. In future studies, higher resolution position estimates, including those from internal Pixhawk sensor fusion, will be used to provide a more consistent comparison across simulation and flight-test.

4.1 Vertical Reposition

In the vertical reposition task, pilots are tasked with a vertical climb from a starting altitude of 6 ft to a target altitude of 14 ft while holding their position. Once the target altitude is achieved, pilots are tasked with maintaining their altitude, position, and heading for 5 seconds, followed by a return to the initial starting altitude of 6 ft. Pilots should hold their position and heading throughout the maneuver within the aforementioned tolerances and complete the maneuver within the allotted time.

4.1.1 Best Runs

To give a sense of performance progression for each Group (A and B) within the study, representative plots of the best runs of the vertical reposition task are superimposed below. Figure 8 shows the vehicle trajectory as viewed from behind throughout the maneuver. This view shows pilot altitude performance and lateral deviation through the maneuver. In both groups, the best participants had at least adequate performance in longitudinal deviation and altitude holding throughout the maneuvers.

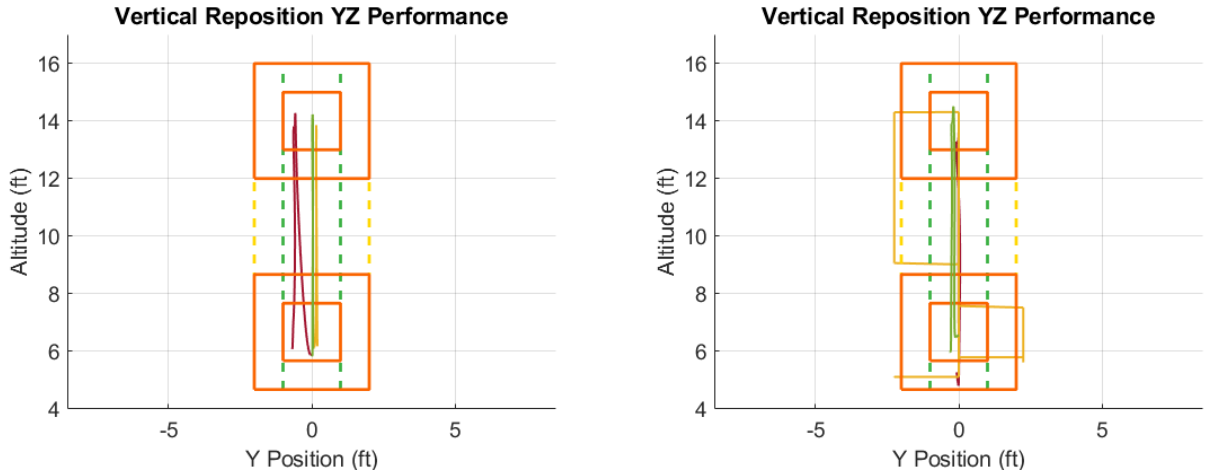


Figure 8 – Best participants in groups A (left) and B (right) for vertical reposition HQTE (altitude)

Figure 8 shows that even in the best runs, performance is notably worse during the flight-test sessions when compared with the equivalent second training session of Group A. While the best pilots in both Group A and Group B could achieve desirable performance in simulation, adequate performance was the best these same pilots could achieve during flight testing.

Figure 9 shows the heading hold performance during vertical reposition tasks for the best Group A and Group B participants. Similar to the results in Fig. 8, the best performance in the flight-test session of Group B is notably worse than the equivalent Group A session in simulation (Fig. 9). However, the degradation is primarily due to a consistent bias in the heading instead of an oscillatory behavior. If the data were centered, the Group B participant would have maintained adequate heading hold performance throughout the maneuver. Once again, the best Group A pilots could achieve desirable heading performance almost immediately in simulation, while even adequate performance was unachievable in flight-test.

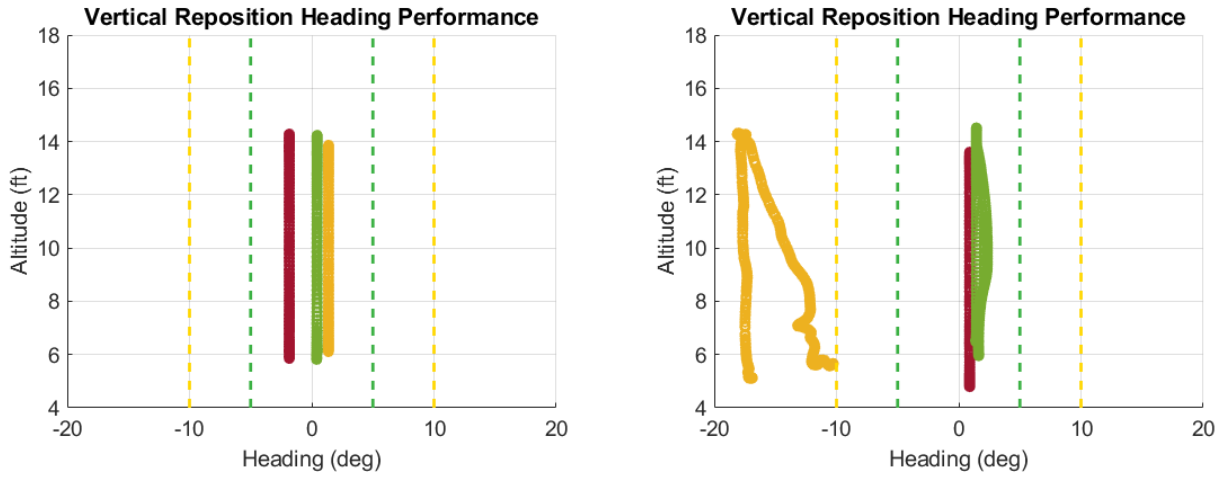


Figure 9 – Best participants in groups A (left) and B (right) for vertical reposition HQTE (heading)

While the best pilots could often achieve desirable performance in the simulation environment within a few practice runs, this was not the case during the flight tests, indicating that there is a higher skill floor

for the flight-tests. This also indicates that task mastery is more difficult (there is a higher skill ceiling) for actual flight. Even the best pilots were unable to overcome the physical differences between the drone handling in simulation vs. flight testing. Despite the differences in training, performance continued to improve throughout the best piloted flight-test runs. Training in flight-test prepared the best pilots just as well as simulation training in the final performance run, indicating that the decreased session two performance did not directly result in diminished performance during the final simulation test.

4.1.2 Stick Activity

The average stick activity RMS values for each of the pilot input axes are shown during the vertical reposition task for session two in Fig. 10 to highlight the differences between simulation and flight-test.

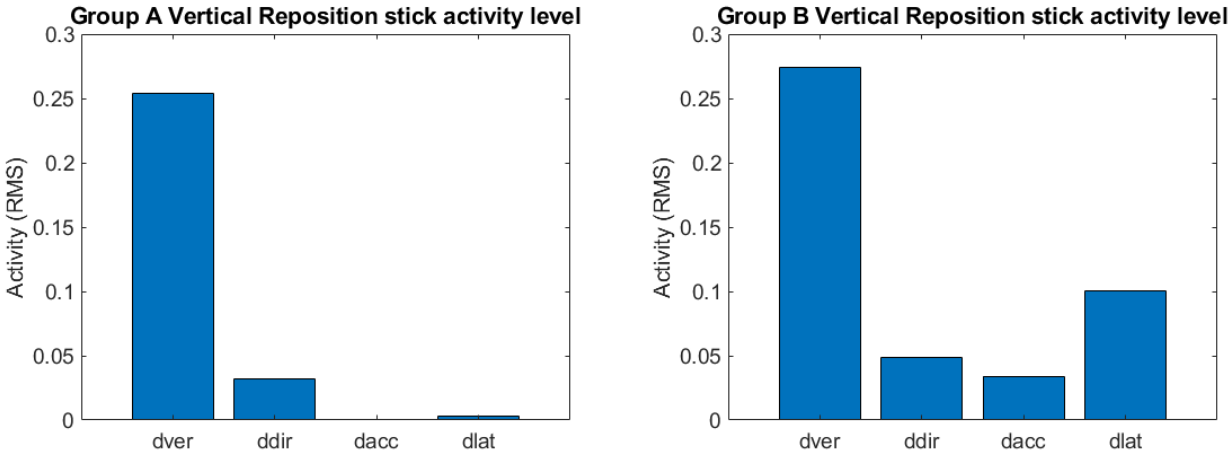


Figure 10 – Vertical reposition average stick activity RMS for session two, all pilots (Group A, Group B)

Intuitively, the vertical axis (δ_{ver}) showed the greatest activity by a factor of at least 2. The critical difference between Group A and Group B, as shown in Fig. 10, was the activity level of the secondary axes δ_{acc}, δ_{lat}, and δ_{dir}. This increased activity is due to pilot action to maintain lateral and longitudinal position and heading. This indicates that the autonomous position hold functionality was not sufficient for pilots to believe they were maintaining desirable position tracking. During simulation, the FCS completely or near completely automates this task for the pilot. Due to substantial differences in the atmospheric conditions of the flight-test, the same FCS could not automate these tasks for the pilot to the desired level, resulting in increased pilot workload and the poorer performance above, even in the best runs. The degradation in position hold performance was also further exacerbated through the use of raw GPS data (which has resolution limitations) for the position hold controller. This results in position “jumps” throughout the flight, which would cause the vehicle to slightly deviate–something pilots are likely to fight.

4.1.3 Performance Progression

The complete performance progression for pilots performing the vertical reposition task is given below. The solid line plots in Fig. 11 and Fig. 12 show pilot performance from one individual run to the next. Simulation sessions are marked with a circle ($\circ$) while flight-test sessions are marked with a diamond ($\diamond$). The dotted lines show the associated second-order curve fit, fitted through each pilot’s first result. Together, these plots show the general trend of each pilot’s performance throughout all three trials. A negatively sloped line indicates increasing performance (reducing error), and a positively sloped line indicates decreasing performance (increasing error).

The Group A learning curve exhibits a noisy exponential decay, while the Group B learning curve is concave (Fig. 11). The concavity is a result of the dramatic increase in maneuver error during flight-test trials. Pilot performance improves in each case on average, demonstrating the training utility of both simulation and flight testing. The same increased difficulty of achieving desirable or adequate results as demonstrated in the case of the best pilots in Fig. 8 and Fig. 9 suggests greater

pilot workload. Decreases throughout the flight-test portion of Group B suggest that, while the skill floor is much higher than in simulation, training is still effective.

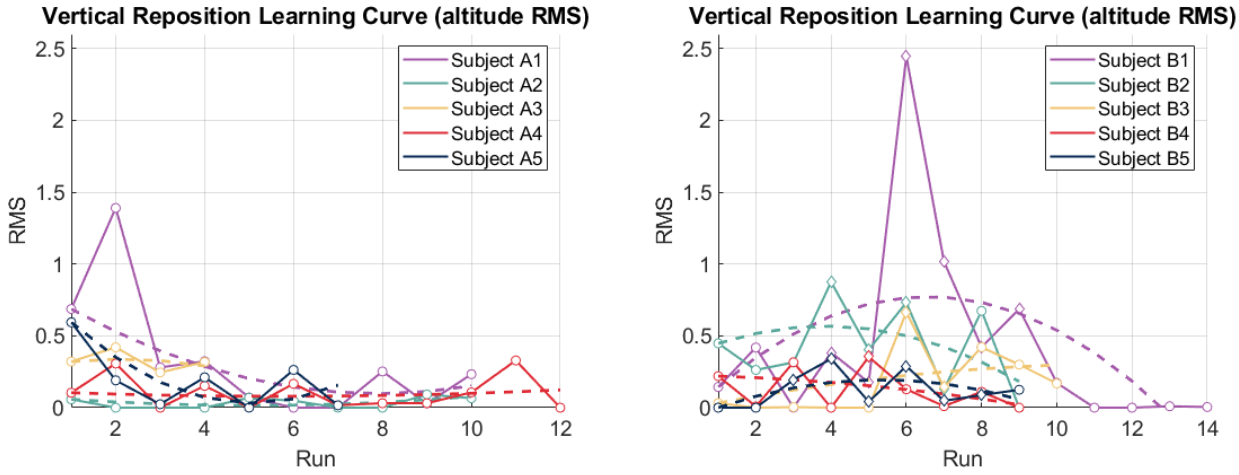


Figure 11 – Vertical reposition altitude learning curves (Group A, Group B) in feet of deviation

As shown in Fig. 12, the Group A learning curve demonstrates consistent performance. In simulation, the FCS autonomously manages the heading. There is essentially nothing for pilots to learn during this task, entailing low pilot workload. A similar but more dramatic decrease in pilot performance in the task is present during the flight-test portion of the training. Performance remains relatively constant, if slightly improving, throughout the flight-test. After returning to simulation, Group B pilots did just as well as those in Group A. The results suggest similar conclusions as Fig. 11; a greater pilot workload, a higher skill floor, and the utility of incorporating subscale flight-test in training.

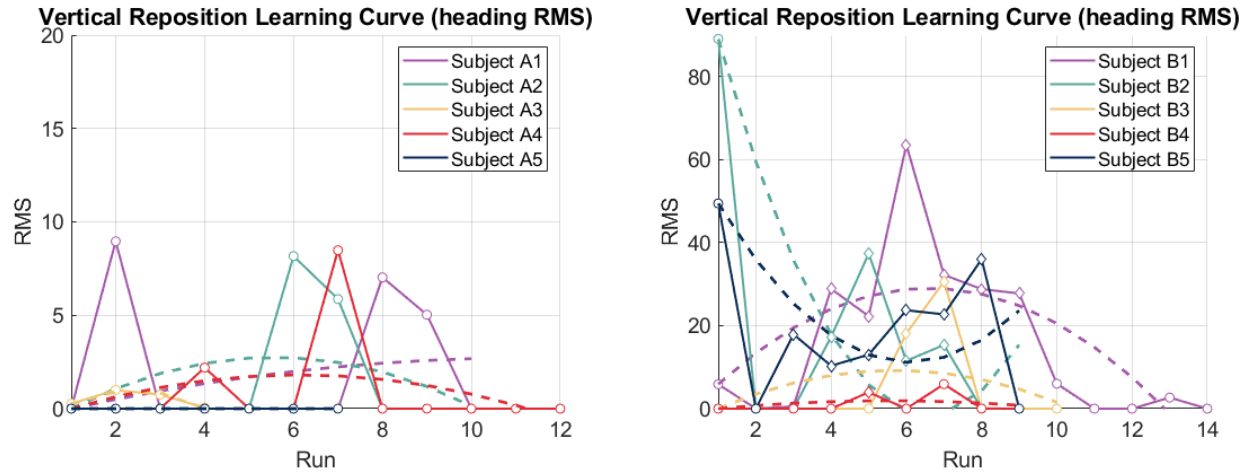


Figure 12 – Vertical reposition heading learning curves (Group A, Group B) in degrees of deviation

4.2 Pirouette

The pirouette task has pilots move centripetally around a central reference, clockwise and counter-clockwise, while maintaining a fixed distance of 33 ft from the reference. Pilots must also maintain their altitude during the maneuver, and complete two revolutions (one in each direction) within the allotted time. Pilot performance is evaluated across their ground track RMS and the difference in their heading from the direction of a line pointing from their current position to the center reference.

4.2.1 Best Runs

To give a sense of performance progression for each Group (A and B) within the study, representative plots of the best runs of the pirouette task are superimposed in Fig. 13. The ground track is plotted as viewed from above, and shows lateral and longitudinal position as the pilot maneuvers around

the central reference point. Figure 14 shows a view of the maneuver projected onto the altitude and heading axes, showing: (i) altitude as the pilot circles the central reference, and (ii) deviation of the true heading from the desired direct heading track over time as the pilot moves along the circle.

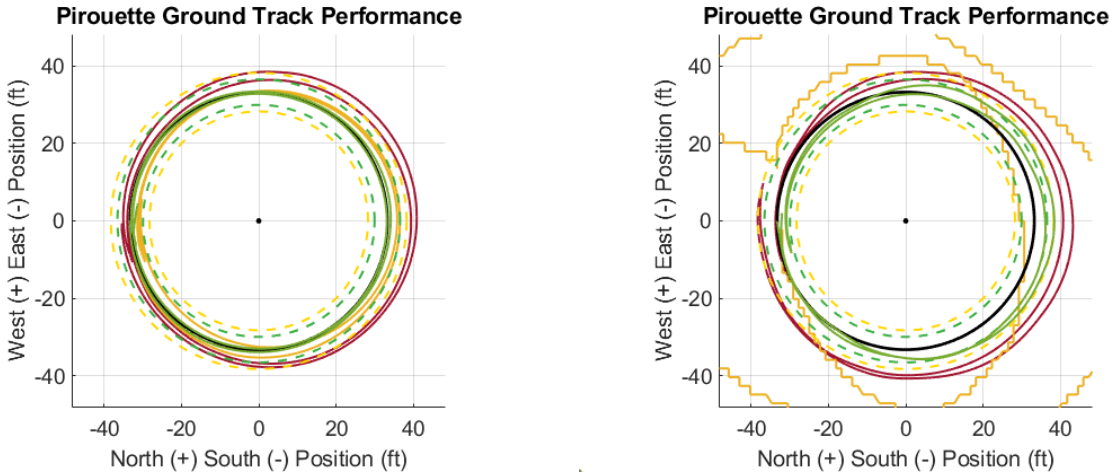


Figure 13 – Best candidates in groups A (left) and B (right) for pirouette HQTE (ground track)

As shown in Fig. 13, performance during the flight-test portion of Group B is significantly worse than the equivalent second training session of Group A. Despite the dramatic differences, the best pilots in Group B showed similar improvements in final performance to those in Group A. In the final runs, while desirable performance was only achieved by the best Group A pilot, the ground tracking errors were similar. The best Group B pilot achieved desirable performance during the vast majority of the final maneuver, illustrating the potential utility of incorporating flight testing in training.

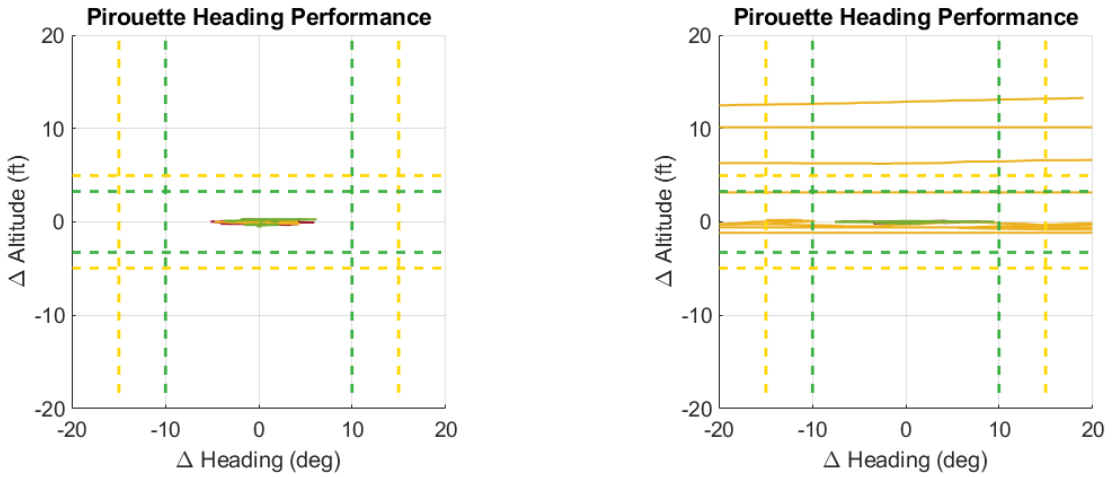


Figure 14 – Best candidates in Group A (left) and B (right) for pirouette HQTE (heading)

Similar results to Fig. 13 are shown in Fig. 14. The best Group A and B initial and final performances were similar, with dramatic differences in the flight-test portion of Group B. These results indicate that both the skill floor and skill ceiling are much higher in the flight-test than in simulation. While desirable results could be achieved in the equivalent second training session of Group A, Group B flight-test results were far from adequate, suggesting significantly greater workload. Despite these differences, as seen in the ground track performance, the best Group B pilots were able to perform as well as Group A pilots in the final performance.

4.2.2 Stick Activity

The average stick activity RMS values for each of the pilot input axes during session two are shown in Fig. 15 to highlight differences between simulation and flight-test.

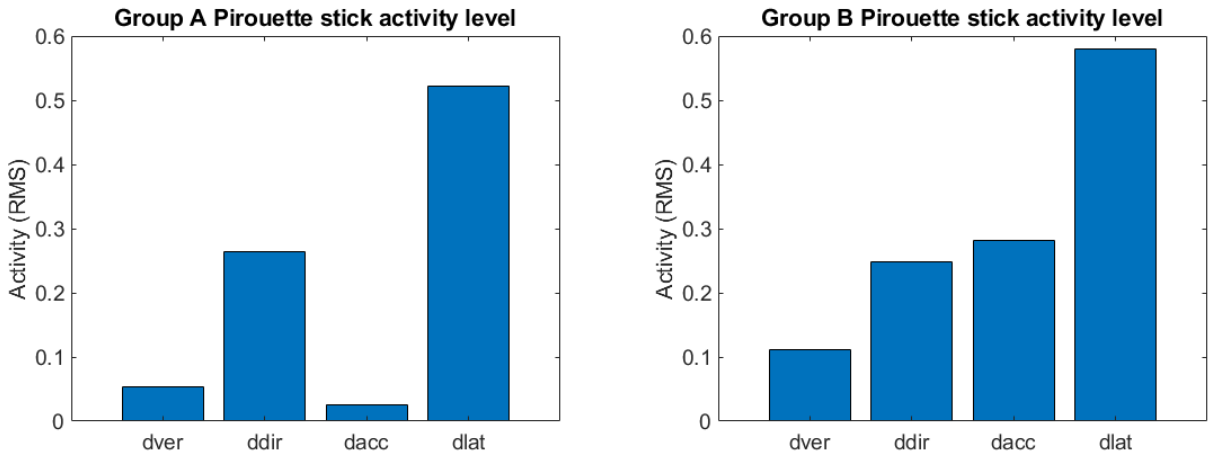


Figure 15 – Pirouette average stick activity RMS for session 2, all pilots (Group A, Group B)

Due to the FCS's altitude hold, the pirouette should be a coordinated maneuver between two primary axes (δ_{lat} and δ_{dir}) as the pilot must move laterally while turning the vehicle's heading towards the center reference. Figure 15 suggests that, while coordination between δ_{lat} and δ_{dir} remained, differences in the longitudinal (δ_{acc}) and vertical (δ_{ver}) axis result from pilot corrections and compensation for external disturbances which were not as severe in simulation. The pilots maneuver off the intended path substantially (as suggested in Fig. 13), resulting in the need for substantial longitudinal corrective action to move the vehicle toward or away from the center reference. This additional stick activity during flight testing suggests greater pilot workload and possibly degraded ergonomics (increased coupling) of the physical inceptor layout compared to the simulation setup.

4.2.3 Performance Progression

The complete performance progression for pilots performing the pirouette task is given in Fig. 16 and Fig. 17. As illustrated in Fig. 11, the solid-line plots show pilot performance from one individual run to the next. The plots are formatted as described in Sec. 4.1.3.

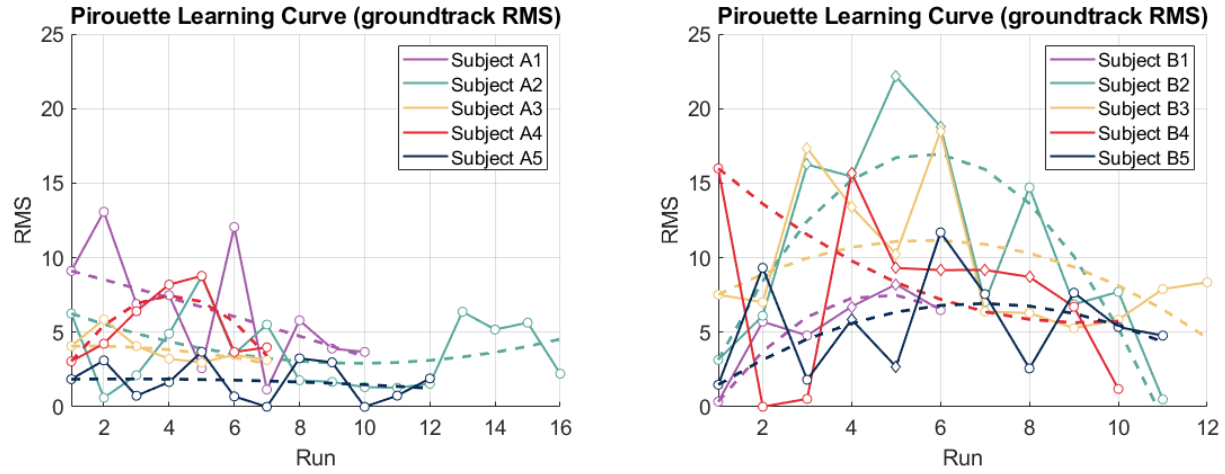


Figure 16 – Pirouette ground track learning curves (Group A, Group B) in feet of deviation

The Group A learning curve in Fig. 16 tends to show decaying improvement, with the exception of subject A2, whose performance at the end inflected, potentially due to increased fatigue due to the abnormal number of repetitions. The Group B learning curve exhibits a similar shape to the Group B learning curves in Fig. 11. On average, performance was significantly worse during Group B's flight-test training. These results suggest greater pilot workload and a higher skill floor during flight testing. Continued performance improvement throughout Group B illustrates the potential utility of incorporating flight testing in training.

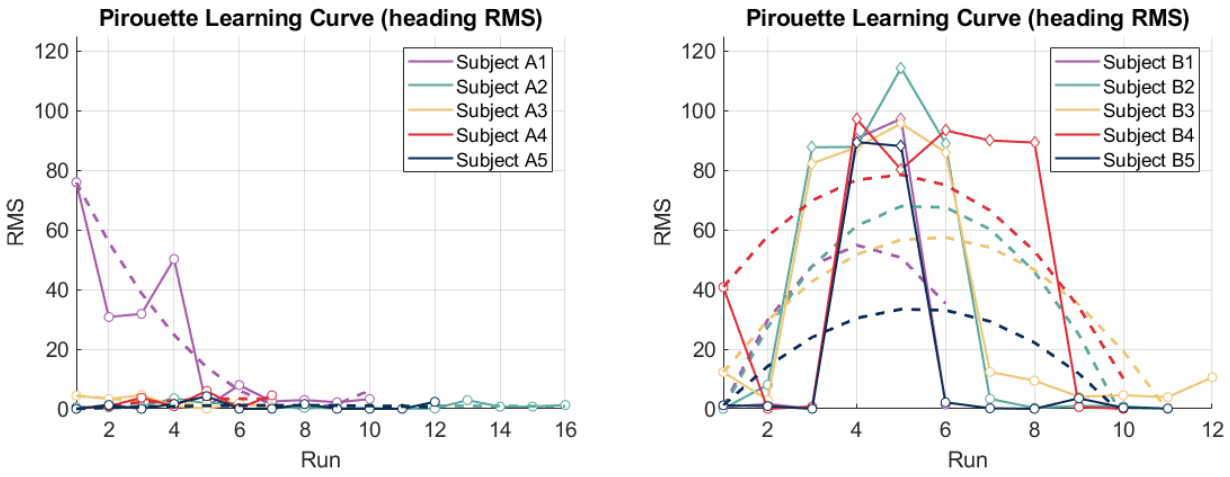


Figure 17 – Pirouette heading learning curve (Group A, Group B) in degrees of deviation

The Group A and Group B learning curves in Fig. 17 share similar characteristics to other Group A and Group B learning curves respectively. Improvement is essentially absent during the flight-test portion of Group B's training. Despite this, performance during the final test was similar. These results suggest much greater pilot workload during the flight-test. Essentially, a task that is impossible to improve in flight-test is easily mastered within a single training session of simulation. This suggests that the skill floor of the task in flight testing is higher than the corresponding skill ceiling in simulation, indicating substantially greater workload in subscale flight testing.

4.3 Turn-and-Hold

The turn-and-hold task begins with pilots at a heading due North with a commanded turn to a heading due West, followed by a commanded turn back toward the initial heading, and vice versa, while maintaining their position and altitude within reasonable allowances. Pilots must also complete the maneuver within the allocated time. For this study, the task was split into two flights performed sequentially. Pilots would perform either the North, West, North sequence or the West, North, West sequence consecutively.

4.3.1 Best Runs

Figure 18 and Fig. 19 show a view of the maneuver projected onto the altitude and heading axes showing altitude as pilots perform the required turns. Discontinuities in the plot are a result of pilot activity in δ_{dir}, which disengages the heading hold. It is inferred that activation of the heading hold implies that the pilot is satisfied with the vehicle's current heading. Figure 20 shows the maneuver projected onto the XY plane, as viewed from above the vehicle, and shows ground track position as the pilot performs the required turns.

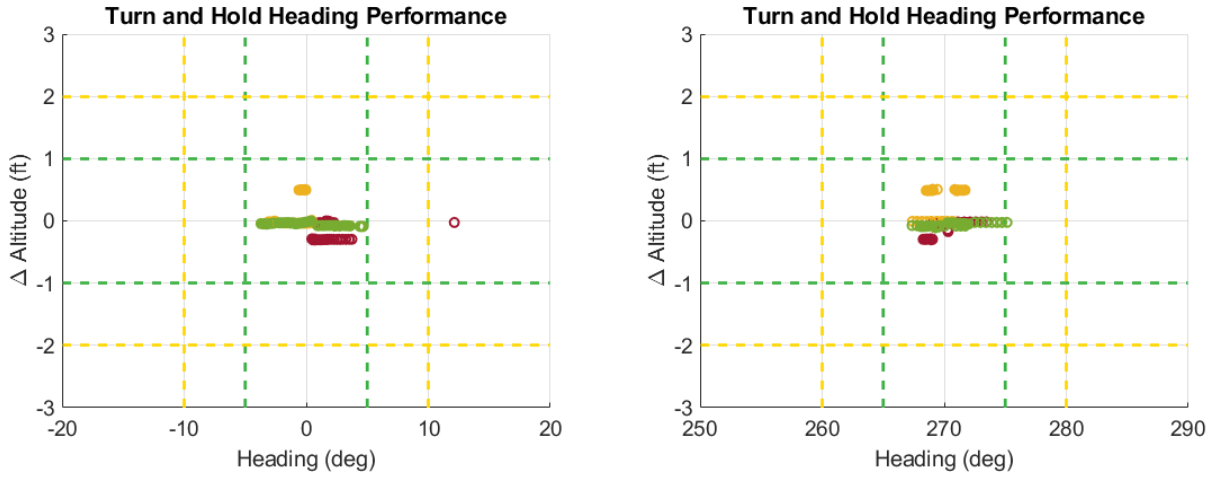


Figure 18 – Best candidate in Group A for turn-and-hold HQTE (heading)

The results in Fig. 18 and Fig. 19 indicate that, similar to Fig. 13, pilots in Group B could not achieve adequate performance in flight-test. This points to a dramatic difference in the skill floor and skill ceiling of the maneuver in simulation vs. in flight-test, suggesting that mastery in simulation may not correspond with even adequate performance in flight testing. Pilots in Group A eventually achieved desirable performance. Group B pilots eventually achieved similar performance, if only adequate. These results show the potential training utility of flight-test training, even when final evaluations are conducted in a simulation environment. These differences may be an artifact of individual pilot skill rather than a result of training differences, as indicated by the averaged progressions later in this section.

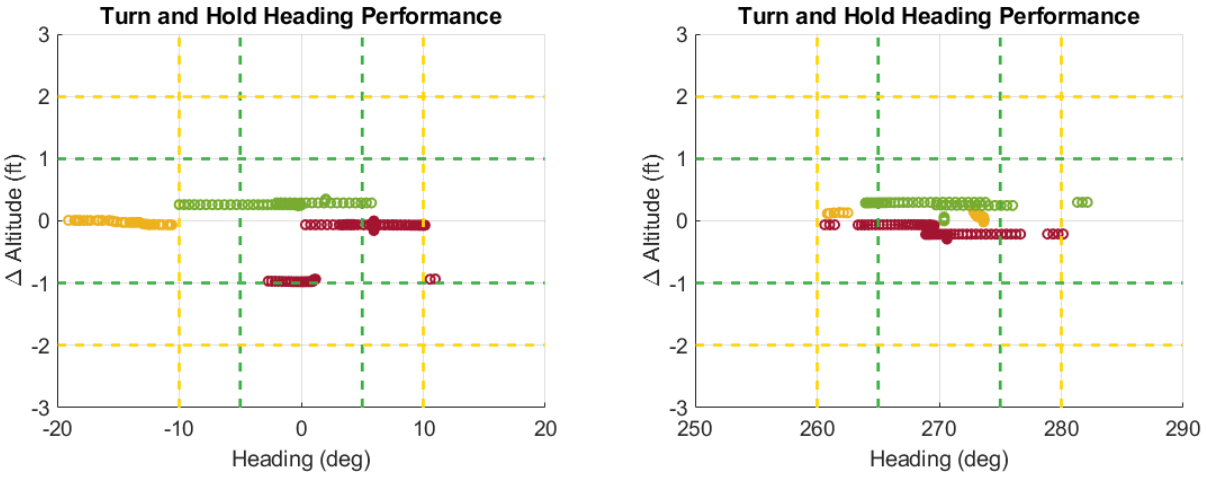


Figure 19 – Best candidate in Group B for turn-and-hold HQTE (heading)

The best piloted performances in Fig. 20 are similar to Fig. 9, with the best Group B pilots achieving desirable performance. While these particular results may suggest that Group B participants performed better in the final test, this may be an artifact of individual pilot skill rather than a result of training differences, as indicated by the average progressions later in this section. Conclusions drawn are similar–but not the same–as those drawn from the results presented in Fig. 17. Quick mastery in simulation does not necessarily indicate the achievement of even adequate performance in flight testing, due to dramatic differences in the skill floor and skill ceiling, respectively.

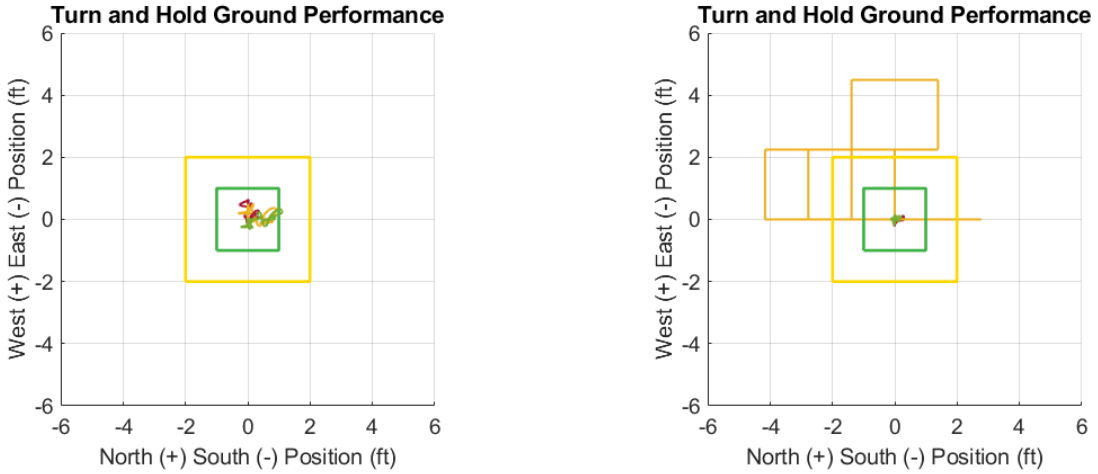


Figure 20 – Best candidates in groups A (left) and B (right) for turn-and-hold HQTE (ground track)

4.3.2 Stick Activity

The average stick activity RMS values for each of the pilot input axes are computed during the turn-and-hold task for all runs in the second session to highlight differences across simulation and flight testing.

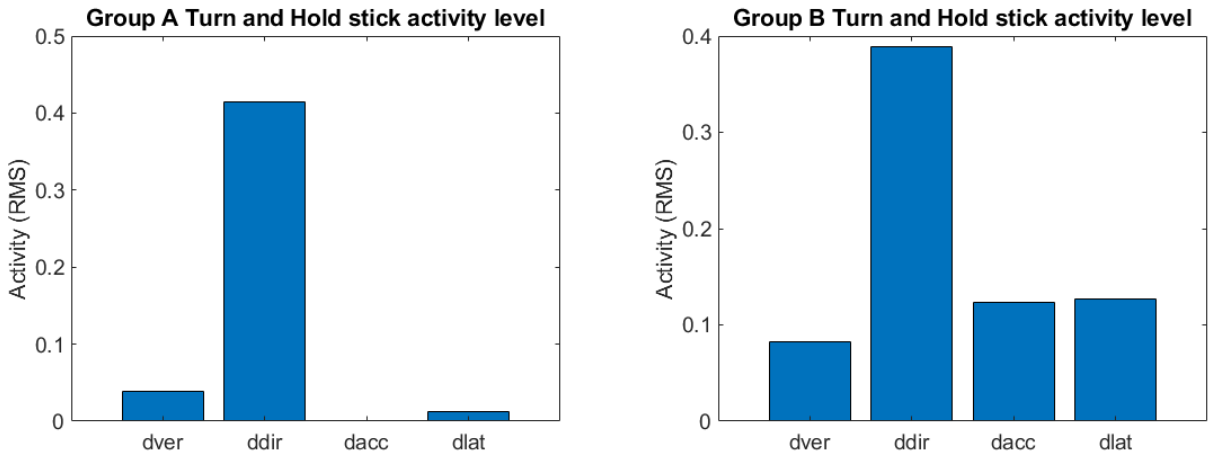


Figure 21 – Turn-and-hold average stick activity RMS for session two, all pilots (Group A, Group B)

The turn-and-hold task is dominated by the directional axis (δ_{dir}), as shown in Fig. 21. In simulation, the small contributions to the other axes are primarily due to minor ergonomic coupling. This is not the case for Group B, likely due to the reduced performance of the position hold during flight testing in the presence of dramatically different atmospheric conditions. This increase in stick activity suggests greater pilot workload during flight testing not present in simulation.

4.3.3 Performance Progression

The complete performance progression for pilots performing the turn-and-hold task are shown below, with solid line plots depicting pilot performance from one individual run to the next and dotted lines indicating the quadratic fittings. The plots are formatted as described in Sec. 4.1.3.

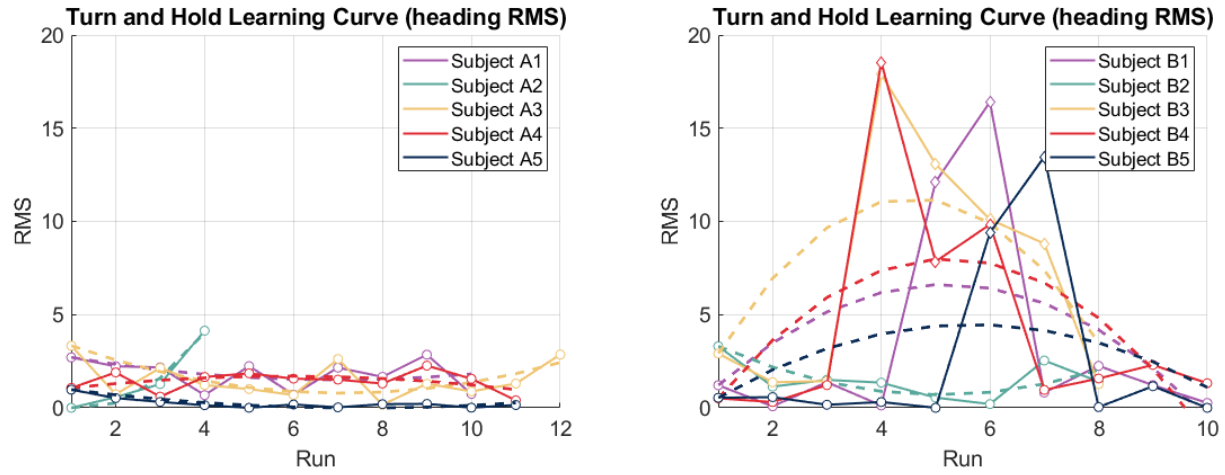


Figure 22 – Turn-and-hold heading learning curves (Group A, Group B) in degrees of deviation

The Group A and B learning curves in Fig. 22 are similar to those in Fig. 12 and suggest similar conclusions. While the FCS does not handle the directional axes autonomously for the pilots, the task is relatively simple (low skill floor) but difficult to master (high skill ceiling). The skill floor dramatically ramps up in flight testing, indicating greater pilot workloads, while training continues to boost improvement. Final performances remain similar across Group A and B, suggesting the potential training utility of flight testing; although, this may be a result of the low skill ceiling in simulation rather than differences in training utility.

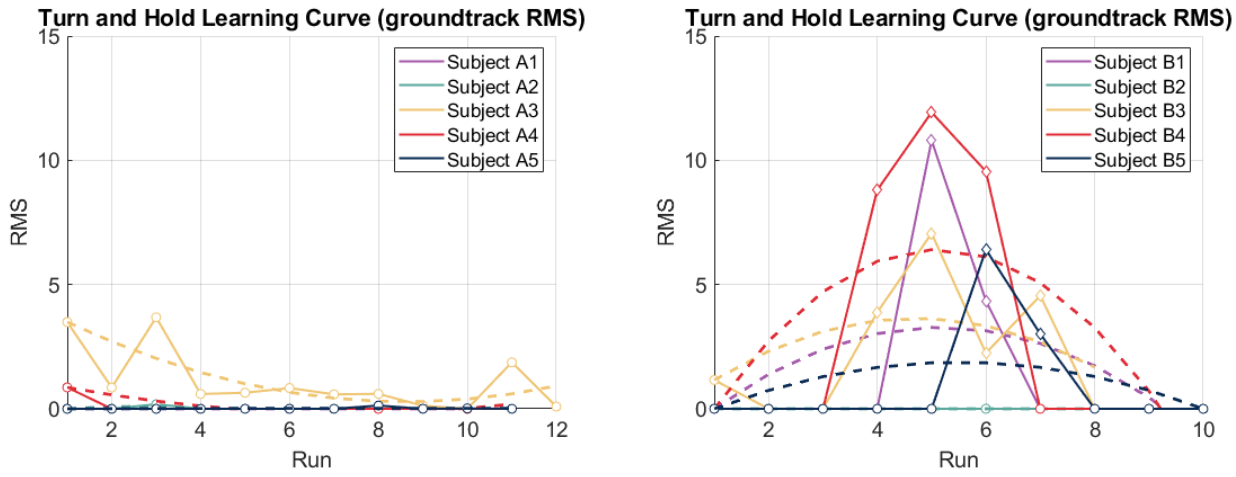


Figure 23 – Turn-and-hold ground track learning curves (Group A, Group B) in feet of deviation

Figure 23 shows that, after initial introduction, Group A and Group B pilots quickly make use of position hold during the turn-and-hold task in simulation. As seen in other performance criteria above, differences in the atmospheric conditions and inceptor ergonomics of the flight-test and simulation may have exacerbated these differences. Worse performance during the flight test portion suggests increased pilot workload throughout the maneuver. As seen in Fig. 17, performance does not improve during the flight-test to a noticeable degree, similarly suggesting that maneuver performance in simulation may not translate to flight-test performance.

4.4 Eye Tracking, Pupil Diameter, and Workload

Heat maps and pupil diameter charts were generated for five pilots in Group A and one pilot in Group B. Data from Group B were insufficient to support conclusions, but they demonstrate the type of analysis enabled by the study design. This is further commented on in the conclusions of this paper. Simulation pilots can be roughly categorized into two groups, those who strongly preferred to use the FPV-style feedback, and those who split their attention between certain instruments (especially the altimeter and compass) and the FPV-style feed. Figure 25 illustrates the heat mapping trends of each view preference.



Figure 24 – Group A eye tracking, FPV-style view preference (left) and split fixation (right)

All simulation pilots used their instrument display to some degree. For certain tasks, the precise feedback from instrumentation was more useful than the rough guide of the reference objects for task completion. The utilization of the altimeter and heading indicator were likely correlated with the nature of the vertical reposition and turn-and-hold courses respectively. Forward and lateral velocities are near zero for the vertical reposition and turn-and-hold tasks, resulting in little to no pilot fixation on these elements during these tasks. For the pirouette course, pilot attention is focused on the FPV-style view, and instruments do not provide immediately useful feedback for maintaining course tracking performance. The proximity of the altimeter and compass to the FPV-style view

may also have increased their utilization. Most pilots frequently glanced but did not fixate on the instrument display during all sessions and maneuvers. It may have been easier to keep the preferred FPV-style view within their peripheral vision while looking at instruments nearer to that view. These findings illustrate the potential utility of instrument displays which allow pilots to retain situational awareness, such as on-screen displays, heads-up displays, and displays that are surrounded by instrument readings.



Figure 25 – Group B eye-tracking in simulation (left) and flight-test (right)

In contrast to Group A, Group B exhibited unique visual attention, as evidenced by gaze behavior. Participant B2 had a gaze concentration similar to that of the Group A participants during simulation training. However, during the flight-test, the fixation appears to be scattered towards the bottom of the displays. This could indicate that the participant frequently fixated on the control sticks. However, during the test session, participant B2 showed gaze patterns similar to those in training session one, with wider fixations than in simulation as degraded performance caused the location of reference objects in their view to widen.

Pupil diameter plots are presented in Appendix A. It should be noted that trial numbers may vary, as turn-and-hold runs were divided into two trials for the sake of pupil diameter measurement. A relative increase or decrease in trial number merely indicates that the participant performed a greater or fewer number of turn-and-hold tasks.

Among Group A participants, pupil diameter was generally below baseline throughout the task. This pattern was predominantly seen in participants A1 and A5, who showed consistently negative values across both training and test sessions, whereas A5 showed the most stable pupil diameter across all trials. In contrast, the pupil diameter of participant A4 fluctuated across trials in all sessions, particularly in training session one, in which trials one, nine, and eleven were above the baseline pupil diameter, possibly indicating elevated cognitive workload. Participant A2 showed the mildest change, with pupil diameter consistently above the baseline in training session one. In comparison to the Group A participants, participant B showed a contrasting pattern where training session two, which was performed on the subscale, showed a decrease in pupil diameter from baseline, while training sessions one and test sessions showed increases. Overall, training session one for participants A2, A4, and B2 indicated a possible high cognitive workload during the session. Pupil diameter alone therefore suggests that, in general, pilot workload decreased as sessions progressed. However, conclusions cannot be drawn from the results for Group B pilots, as the authors were unable to establish a baseline in the flight-test environment and had to use the baseline from the simulation session (session 1). The authors hypothesize that, had a baseline been performed for the B2 participant, the pupil diameter would have shown a noticeable increase from the baseline due to the increased stress levels during actual flight testing.

4.5 Subjective Assessments

To highlight differences in training utility across each group, this section summarizes the trends in self reported handling qualities and workload through the Cooper-Harper and NASA-TLX surveys conducted at the end of each session. Notable differences across these measures of perceived

improvement are discussed. This section also summarizes the relationship between past experiences with FPV drones and flight simulators and the resulting quantitative performance measures.

4.5.1 Cooper-Harper, NASA-TLX, and Training Utility

Figure 26 shows each pilot's progressive Cooper-Harper handling qualities rating after each session. These ratings allow pilots to quantify the relative ease or difficulty with which they controlled the vehicle throughout the session. The Cooper-Harper handling qualities rating scale workflow is presented in Appendix C.

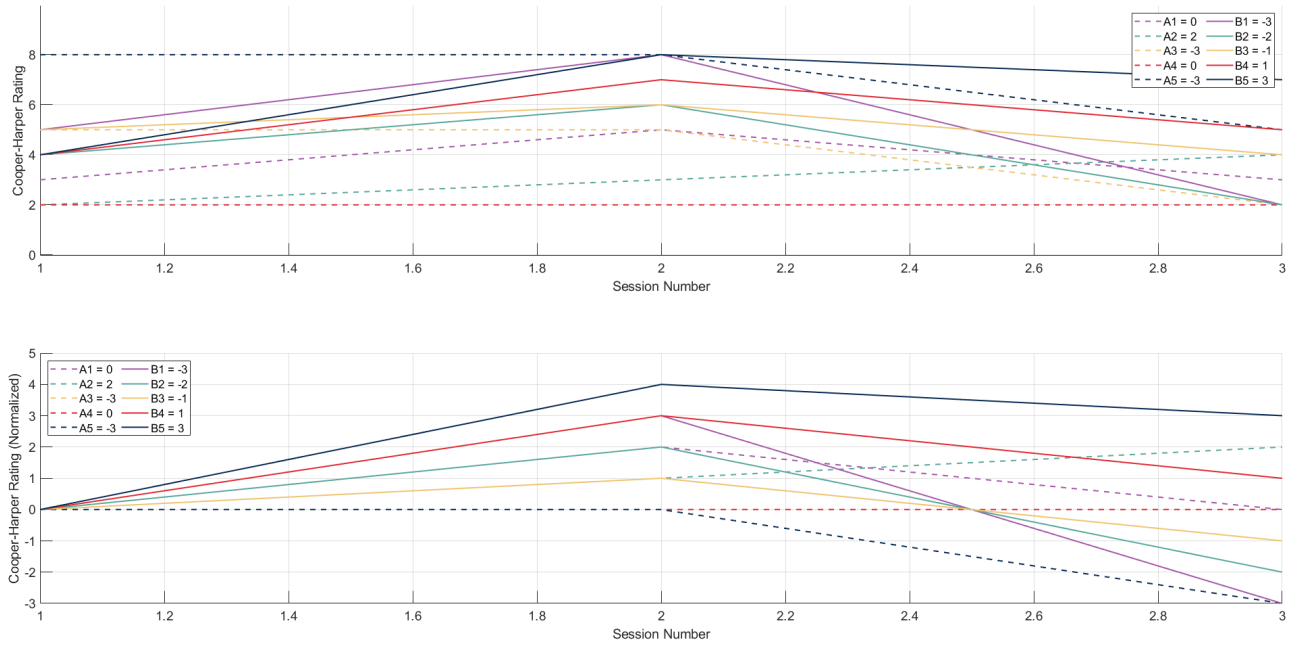


Figure 26 – Cooper-Harper ratings in groups A (dashed) and B (solid)

Because training is evaluated in terms of relative improvement, Fig. 26 shows both raw and normalized Cooper-Harper ratings. Considering the normalized rating trends, two Group A and three Group B pilots reported improvement. Across both groups, pilots did not ever report an improvement in handling qualities until the last training session. All Group B pilots felt that handling qualities had improved following a return to simulation from flight testing, whereas some Group A pilots reported worsening handling qualities after all sessions were complete. Incorporating flight testing may have improved the training experience by boosting pilot perception of their performance in simulation by connecting their experience with real flight.

Figure 27 shows surveyed NASA-TLX workload responses collected after each pilot completed a training or test session. As training is a measure of relative improvement, Fig. 28 shows these survey responses normalized to each pilot's first response. When interpreting the results from the NASA-TLX workload responses, the subfigures represent the ratings in response to the following categories:

- **Mental:** How mentally demanding was the task?
- **Physical:** How physically demanding was the task?
- **Temporal:** How hurried or rushed was the pace of the task?
- **Performance:** How successful were you in accomplishing what you were asked to do?
- **Effort:** How hard did you have to work to accomplish your level of performance?
- **Frustration:** How insecure, discouraged, irritated, stressed, and annoyed were you?

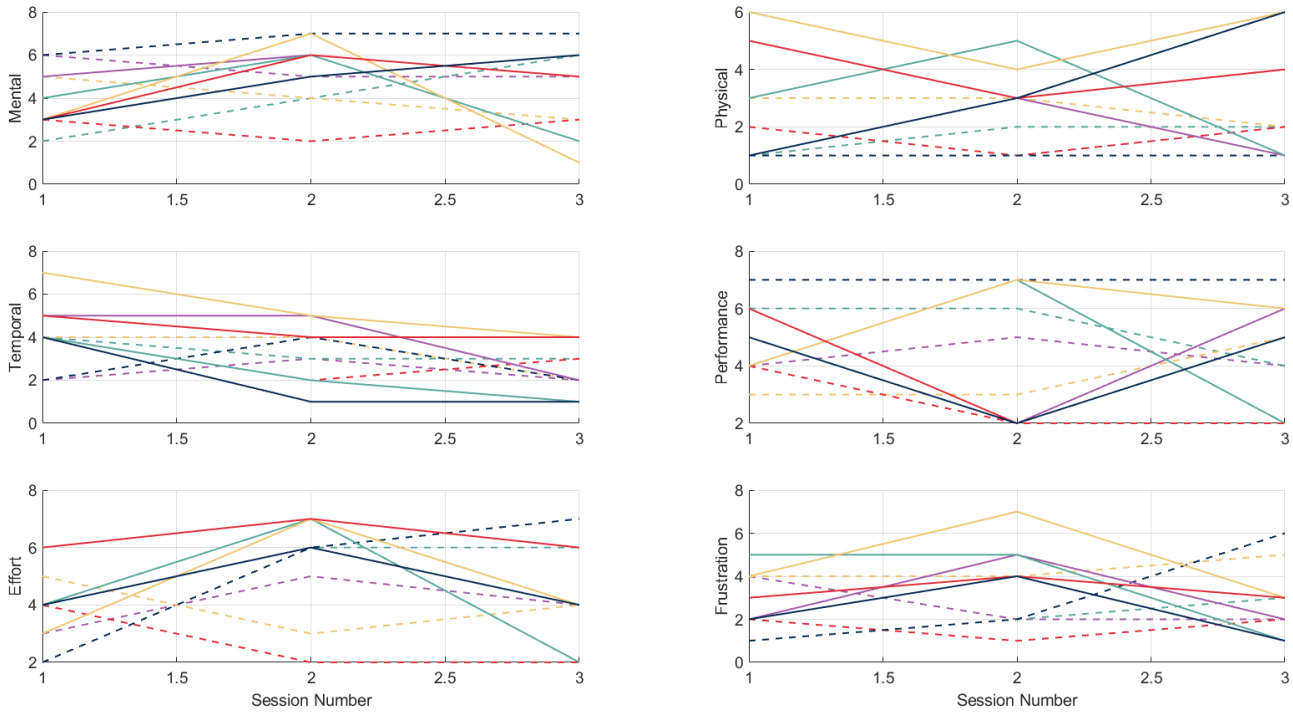


Figure 27 – NASA-TLX ratings in groups A (dashed) and B (solid)

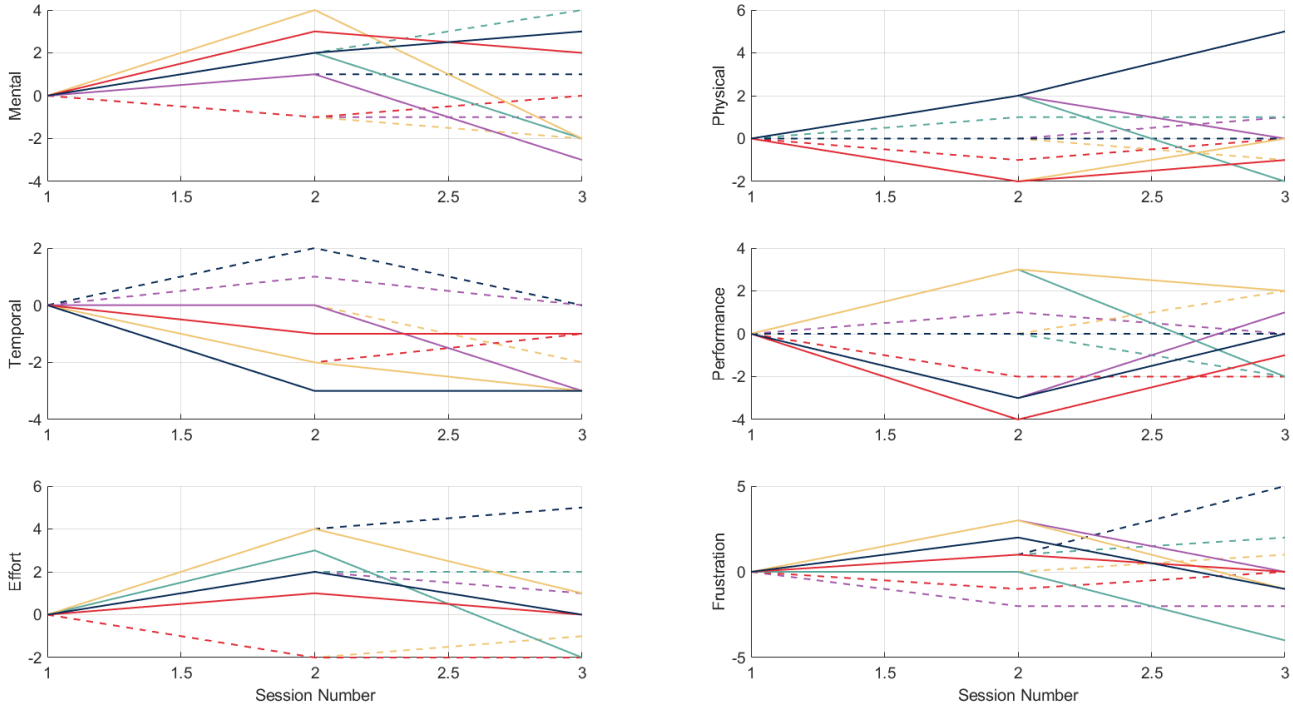


Figure 28 – NASA-TLX normalized ratings groups A (dashed) and B (solid)

The NASA-TLX ratings in Fig. 28 suggest that exposure to flight-test in Group B made pilots more likely to perceive improvement in mental, physical, performance, and temporal demands, as well as in managing the frustrations of flight in their final performance. In contrast, Group A pilots were more likely to report a decrease in the effort demand for the final performance. The long duration of Group B's training sessions, considering the time to travel to, setup, and perform the flight-test, may have contributed to increased fatigue and perceived effort demand. Flight testing, by connecting pilots' training with real flight, may have boosted morale and workload demand perception, as was concluded from Fig. 26.

4.5.2 Prior Experience and Performance

In Group A, as seen in the learning curves above, both the best (A5) and worst (A1) performers were from pilots without FPV drone or flight simulator experience. All other pilots, both with (A3, A4) and without (A2) flight simulator or FPV drone experience, were equally likely to perform as well as the others. In Group B pilot B5, who did not have any flight simulator or FPV drone experience, consistently performed better than the other pilots. Among the remaining pilots, those with (B1, B2, B4) and without (B3) were equally likely to perform as well as the others.

The results from both groups do not suggest any change in performance associated with prior experience with FPV drones or flight simulators. This is potentially due to the utilization of a flight control system developed around SVO, which emphasizes accessible and intuitive control of the vehicle for pilots regardless of experience [12, 14]. Future work could explore how similarity in control system familiarity from past experiences influences performance, and how the simplicity of the control system affects performance.

4.5.3 Training Utility

The overwhelming majority of pilots in both Group A and Group B positively assessed their flight time as a training exercise. Only some Group B pilots presented this assessment with caveats, emphasizing the difficulty of real flight and the necessity for additional training. One Group B pilot commented on discrepancies between the control actuators in simulation and flight-test, indicating that coupling was more common in flight-test, although another Group B pilot highlighted discrepancies in the atmospheric conditions across simulation and flight testing. Group A pilots also mentioned struggling with coupling just in simulation. Across each group, pilots commented on high workload and stress. Future work should improve consistency across atmosphere, inceptor ergonomics, and flight dynamics models. These improvements could better illustrate the differences between training progression across simulation and flight testing in an unbiased manner.

4.6 Conclusions

Pilot performance across all tasks was objectively worse in flight testing than simulation. When the best Group A pilots achieved desirable performance in simulation, the best that any flight-testing Group B pilot could achieve was adequate performance. In certain tasks, such as pirouette and turn-and-hold, desirable performance in simulation could not be matched by adequate performance in flight testing. These differences should be considered with the understanding that: (i) atmospheric conditions modeled in simulation were significantly less severe than those present during flight testing, (ii) the inceptor setup in flight testing was perceived to have more cross-axis coupling, and (iii) the simulation model may have had different response characteristics. With these caveats, pilots improved their performance in four measures during the flight-test portion of their training, suggesting equal objective utility in training that incorporates flight testing and simulation only training.

Stick activity measurements across all tasks suggest greater pilot workload during flight testing. Group B pilots averaged substantially more activity on secondary axes during flight testing than Group A pilots in their corresponding second training session. Increased stick activity may have been due to ergonomic inconsistencies mentioned by some Group B pilots. Future work will reduce ergonomic inconsistencies between simulation and flight testing, likely targeting the use of hand-held USB transmitters which do not require customized mounting solutions.

Eye-tracking and pupil diameter data during simulation indicated decreasing workload as training sessions progressed, suggesting that Group A pilots learned to adapt to the demands of simulated flight. Eye-tracking results also indicate the preference for some pilots to split their attention between an instrument display and the FPV-style view. Eye-tracking and pupil diameter data were not collected in enough abundance for conclusions to be drawn about Group B participants; though, the authors hypothesize that the stress levels would have been increased during flight testing. Future work should expand on the link between eye-tracking, pupil diameter, heart rate activity, and pilot workload in training that incorporates simulation and flight testing, as well as the utility of instrument displays alongside FPV-style views.

Subjective workload information, as well as background profiles on previous flight experiences, was collected through surveys. Past experience with FPV drones and flight simulators was not directly

shown to be associated with better performance, potentially due to the selection of an FCS developed around SVO. Future work should explore the relationship between the simplicity of the control system and objective performance progression. The trends in subjective assessments of Cooper-Harper handling qualities and NASA-TLX workload indicate that pilots whose training incorporated flight testing were more likely than those who trained only in simulation to perceive that they improved throughout the training sessions–at the expense of higher effort. Incorporating flight testing into piloted training may have improved perceived handling-qualities and workload assessments by improving pilot morale and connecting their experience with real flight. Put another way, the severe degradation in performance during the flight-testing sessions may have led to a feeling of greater improvement when Group B participants began their test sessions.

While the objective performance improvement of pilots whose training incorporated flight testing was equal to that of pilots who trained in simulation only, there are notable differences in perceived training utility and pilot workload. Pilots were more likely to report reductions in workload and improvements in handling qualities in their final simulation session if their training incorporated flight testing. As a preliminary paper, the results indicate that there is a potential benefit to incorporating subscale flight testing into the pilot training workflow. Future studies will incorporate suggested improvements to further refine these findings.

5. Acknowledgments

This research was partially funded by the US Economic Development Administration Grant 08-69-05693, the George Kaiser Family Foundation, and Tulsa Innovation Labs. This research was performed under IRB-26-218. The authors would like to thank Sarena Beatriz Seno, Steven Yon, Jamie Lam, and the members of the Simulation to Flight Applied Research Laboratory for their assistance in the development of the HQTE courses.

6. Contact Author Email Address

The corresponding author for this paper is Antonio Chappell, tony.chappell@okstate.edu, Cell: (405) 818-7713 | Address: 526 N Elgin Ave, Tulsa, OK, USA 74106.

7. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] S. J. Hamstra, R. Brydges, R. Hatala, B. Zendejas, and D. A. Cook. Reconsidering Fidelity in Simulation-Based Training. *Academic Medicine*, 89(3):387–392, 2014.

[2] J. P. Fuentes-García, V. J. Clemente-Suárez, M. Á. Marazuela-Martínez, J. F. Tornero-Aguilera, and S. Villafaina. Impact of Real and Simulated Flights on Psychophysiological Response of Military Pilots. *International Journal of Environmental Research and Public Health*, 18(2):787, 2021.

[3] P. E. O'Connor, C. K. McClernon, J. S. Warm, et al. Stress Training Improves Performance During a Stressful Flight. *Human Factors*, 53(3):207–218, 2011.

[4] C. K. McClernon, M. E. McCauley, P. E. O'Connor, and J. S. Warm. Stress Training Enhances Pilot Performance During a Stressful Flying Task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 1922–1926, 2010.

[5] Z. Gomolka, D. Kordos, and E. Zeslawska. The Application of Flexible Areas of Interest to Pilot Mobile Eye Tracking. *Sensors*, 20(4):986, 2020.

[6] J. H. Park, L. Chen, I. Higgins, Z. Zheng, S. Mehrotra, K. Salubre, M. Mousaei, S. Willits, B. Levedahl, T. Buker, et al. How is the Pilot Doing: VTOL Pilot Workload Estimation by Multimodal Machine Learning on Psycho-physiological Signals. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 2311–2318. IEEE, 2024.

[7] A. Sobron, D. Lundström, and P. Krus. A Review of Current Research in Subscale Flight Testing and Analysis of Its Main Practical Challenges. *Aerospace*, 8(3):74, 2021.

[8] C. H. Wolowicz, J. S. Bowman, and W. P. Gilbert. Similitude Requirements and Scaling Relationships as Applied to Model Testing. NASA Technical Paper NASA TP-1435, National Aeronautics and Space Administration, August 1979.

[9] C. M. Ivler, K. Russell, A. Gong, T. Berger, and M. J. S. Lopez. Toward a UAS Handling Qualities Specification: Development of UAS-Specific MTEs. In *Vertical Flight Society 78th Annual Forum & Technology Display*, Fort Worth, TX, USA, May 2022.

[10] D. Klyde, D. Mitchell, M. Shubert, and M. Jones. Handling Qualities Test Guide for Powered Lift VTOL Capable Aircraft with Indirect Flight Controls and Operating in Day Time Visual Flight Rules. Technical Report DOT/FAA/TC-23/59, Federal Aviation Administration, William J. Hughes Technical Center, Aviation Research Division, March 2024. Final Report. Includes an HQTE catalog with task descriptions (e.g., pirouette) and performance requirements.

[11] A. Comer, Z. Atkinson, P. Bhandari, Z. Miller, and Harp E. Development of a Simulation to Flight Workflow for Subscale Flight Testing of Experimental Control Laws. In *AIAA SCITECH 2026 Forum*, Orlando, FL, January 2026. AIAA-2026-2063.

[12] A. Comer, I. Chakraborty, S. Putra, R. Bhandari, and B. Kunwar. Flight Testing a Trajectory Control System on a Subscale Transitioning VTOL Aircraft. *Journal of Guidance, Control, and Dynamics*, 2025.

[13] A. Comer. *A Generalized Flight Control Architecture for Transitioning Flight Vehicles with Flight Test Validation*. PhD thesis, Auburn University, Auburn, AL, USA, December 2024.

[14] I. Chakraborty, A. Comer, R. Bhandari, S.H. Putra, A.A. Mishra, R. Schaller, D. Sizoo, and R. McGuire. Piloted Simulation-Based Assessment of Simplified Vehicle Operations for Urban Air Mobility. *AIAA Journal of Aerospace Information Systems*, 21(5), 2024.

[15] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[16] R. C. Sarker, A. Pudasaini, P. Bhandari, Z. Atkinson, A. Comer, N. Fardin, P. H. Dabaghian, and A. Halder. Design, development, and subscale flight testing of an emergency response evtol aircraft. In *Vertical Flight Society Vertical Lift Aircraft Design and Aeromechanics Specialist's Conference*, San Jose, California, USA, January 2026. Vertical Flight Society. Presented January 27–29, 2026.

[17] D. Klyde, D. Mitchell, M. Shubert, and M. Jones. Handling qualities test guide for powered lift vtol capable aircraft with indirect flight controls and operating in day time visual flight rules. Technical Report DOT/FAA/TC-23/59, U.S. Department of Transportation, Federal Aviation Administration, Technical Center, Atlantic City, NJ, USA, 2023. Systems Technology, Inc. Report STI-TR-1484-1.

[18] AirportsBase. Tulsa, oklahoma airport listings. `https://airportsbase.org/United_States/Oklahoma/Tulsa`, n.d. Accessed: 2026-06-13.

[19] X-Plane Development Team. Xplane2blender: Scenery and aircraft export addon for blender and x-plane. `https://github.com/X-Plane/XPlane2Blender`, 2025. GitHub repository, accessed June 10, 2026.

[20] I. Chakraborty, A. M. Comer, R. Bhandari, S. H. Putra, A. A. Mishra, R. Schaller, D. Sizoo, and R. McGuire. Piloted simulation-based assessment of simplified vehicle operations for urban air mobility. *Journal of Aerospace Information Systems*, 21(4):392–421, 2024.

[21] A. Comer, Z. Atkinson, P. Bhandari, and Z. Miller. Development of a Simulation to Flight Workflow for Subscale Flight Testing of Experimental Control Laws. In *AIAA SCITECH 2026 Forum*, Orlando, FL, June 2026. (To Appear).

[22] A. Giorgi, V. Ronca, A. Vozzi, N. Sciaraffa, A. Di Florio, L. Tamborra, A. Simonetti, P. Aricò, G. Di Flumeri, and D. Rossi. Wearable Technologies for Mental Workload, Stress, and Emotional State Assessment During Working-Like Tasks: A Comparison with Laboratory Technologies. *Sensors*, 21(7):2332, 2021.

[23] A. Hongn, F. Bosch, L. E. Prado, J. M. Ferrández, and M. P. Bonomini. Wearable Physiological Signals under Acute Stress and Exercise Conditions. *Scientific Data*, 12:520, 2025.

[24] P. Saravanan, T. Nisar, Q. Zhang, F. Masud, and F. Sasangohar. Occupational stress and burnout among intensive care unit nurses during the pandemic: A prospective longitudinal study of nurses in covid and non-covid units. *Frontiers in Psychiatry*, 14:1129268, 2023.

[25] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma. A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. *Sensors*, 23(7):3565, 2023.

[26] L. Cheng, Y. C. Shen, Q. He, and M. J. Zhang. Spying with a Pilot's Eye: Using Eye Tracking to Investigate Pilots' Attention Allocation and Workload During Helicopter Autorotative Gliding. *Heliyon*, 10(16):e35872,

2024.

[27] J. Greiwe and N. Friedrich. Gaze Movements of Helicopter Pilots During Real and Simulated Take-Off and Landing Maneuvers. *Aerospace*, 11(6):429, 2024.

[28] U. S. Karunathilake, C. Meissner, and P. Saravanan. Eye-tracking metrics and visuospatial task complexity: A comparative analysis of mental workload indicators. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 69(1):2188–2195, 2025.

[29] Tobii. Tobii pro glasses 3 (specifications and product information). `https://www.tobii.com/products/eye-trackers/wearables/tobii-pro-glasses-3`. Accessed: 2026-01-08.

[30] S. Mathôt. Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition*, 1(1):16, 2018.

[31] R. P. Van Der Wel and H. Van Steenbergen. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, 25(6):2005–2015, 2018.

[32] C. Kelbsch, T. Strasser, Y. Chen, B. Feigl, P. D. Gamlin, R. Kardon, T. Peters, K. A. Roecklein, S. R. Steinhauer, E. Szabadi, A. J. Zele, H. Wilhelm, and B. J. Wilhelm. Standards in Pupillography. *Frontiers in Neurology*, 10:129, 2019.

[33] G. E. Cooper and R. P. Harper. The use of pilot rating in the evaluation of aircraft handling qualities. NASA Technical Note NASA TN D-5153, National Aeronautics and Space Administration, 1969.

[34] G. Lintern and T. Hughes. Development of a supervisory control rating scale. Technical Report AFRL-RH-WP-TR-2008-0054, Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio, January 2008.

A Pupil Diameter Plots

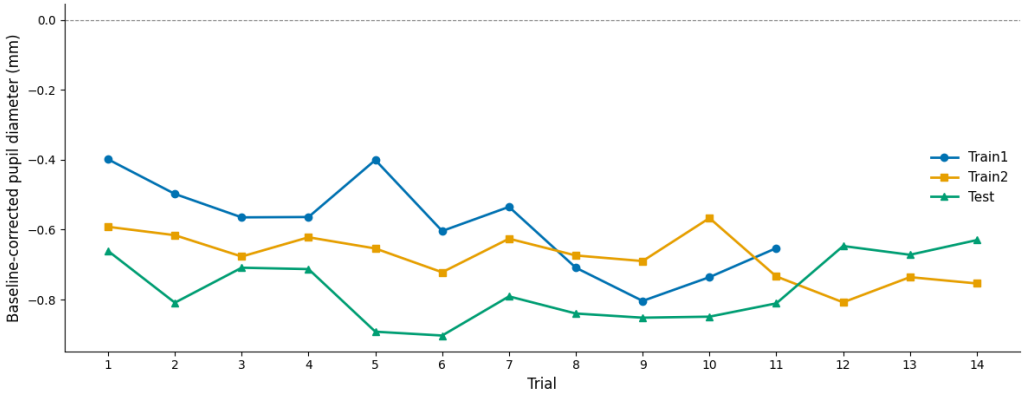


Figure 29 – Participant A1 pupil diameter

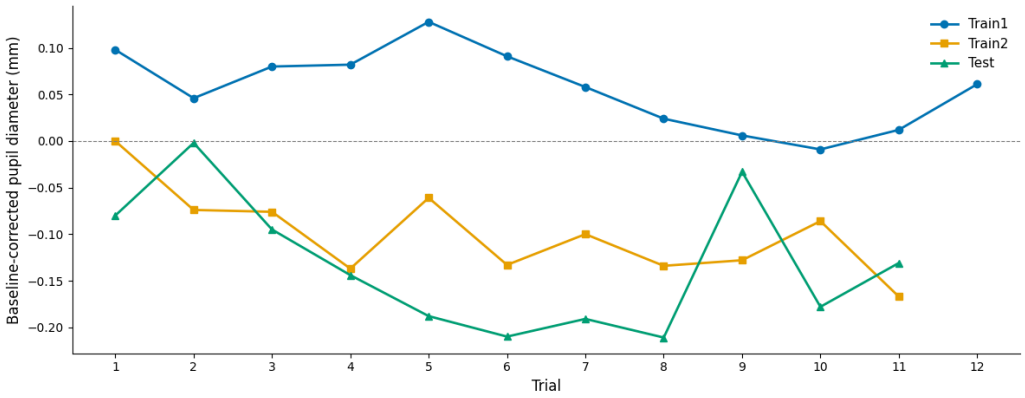


Figure 30 – Participant A2 pupil diameter

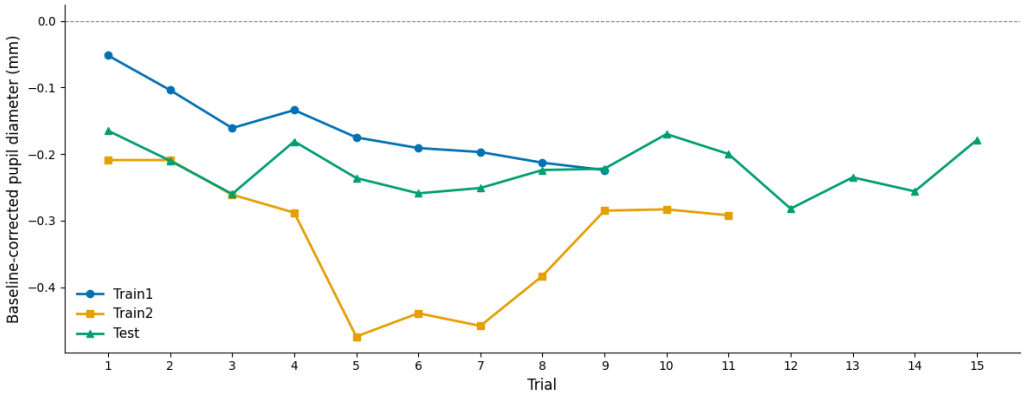


Figure 31 – Participant A3 pupil diameter

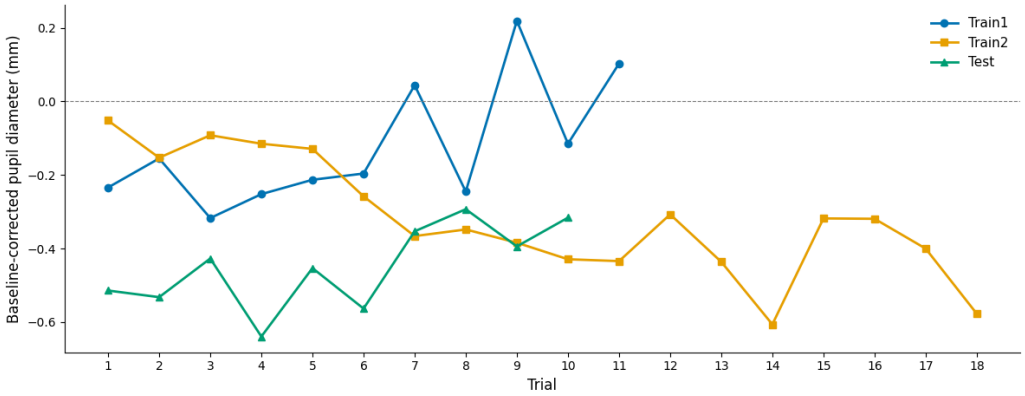


Figure 32 – Participant A4 pupil diameter

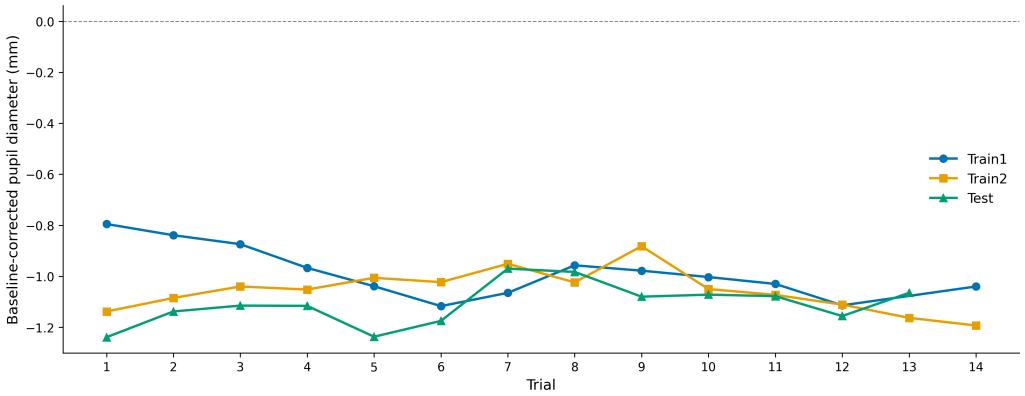


Figure 33 – Participant A5 pupil diameter

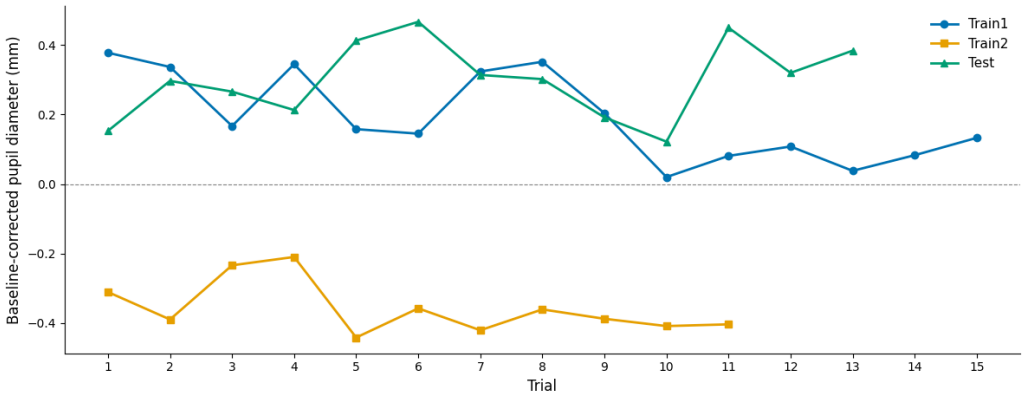


Figure 34 – Participant B2 pupil diameter

B Pilot Setups

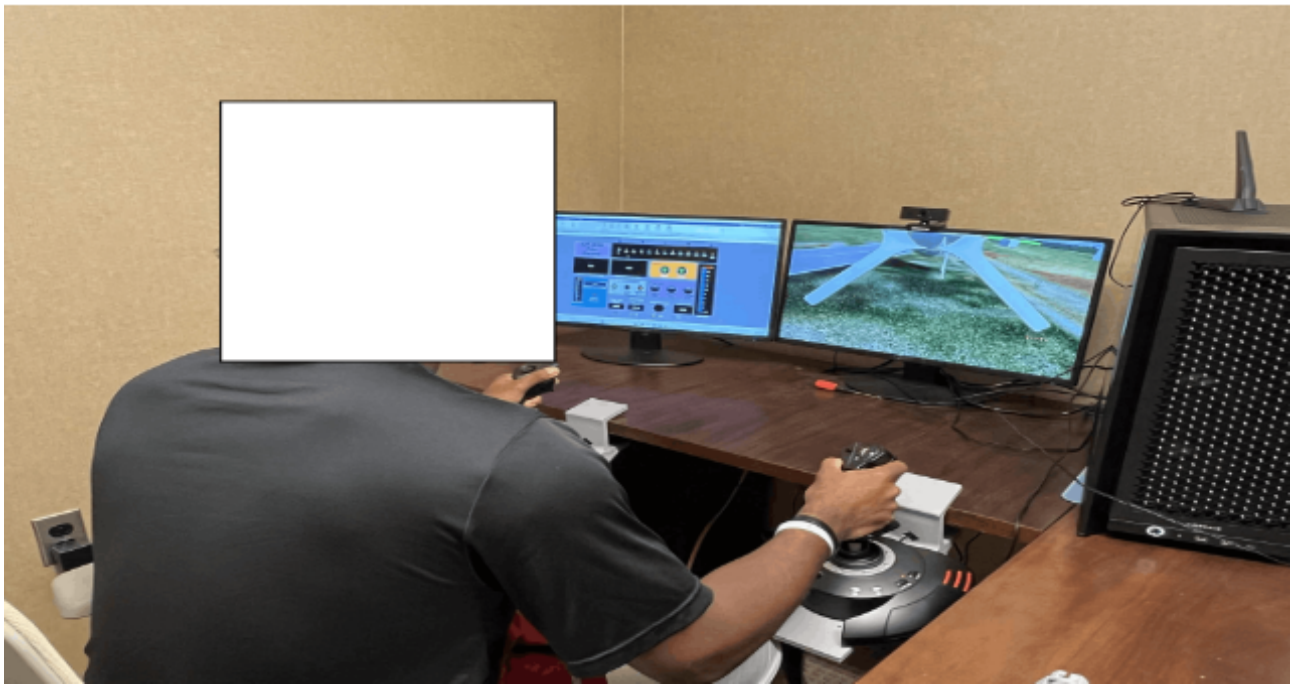


Figure 35 – Simulation setup (simulation)

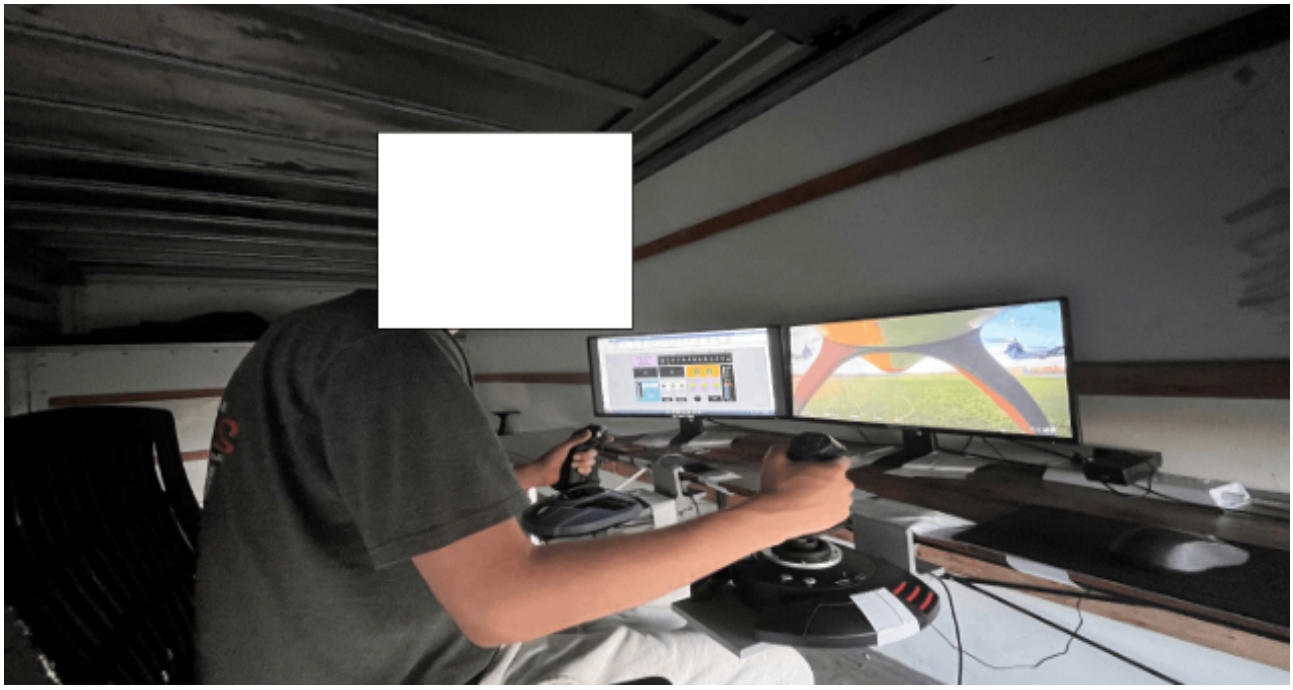
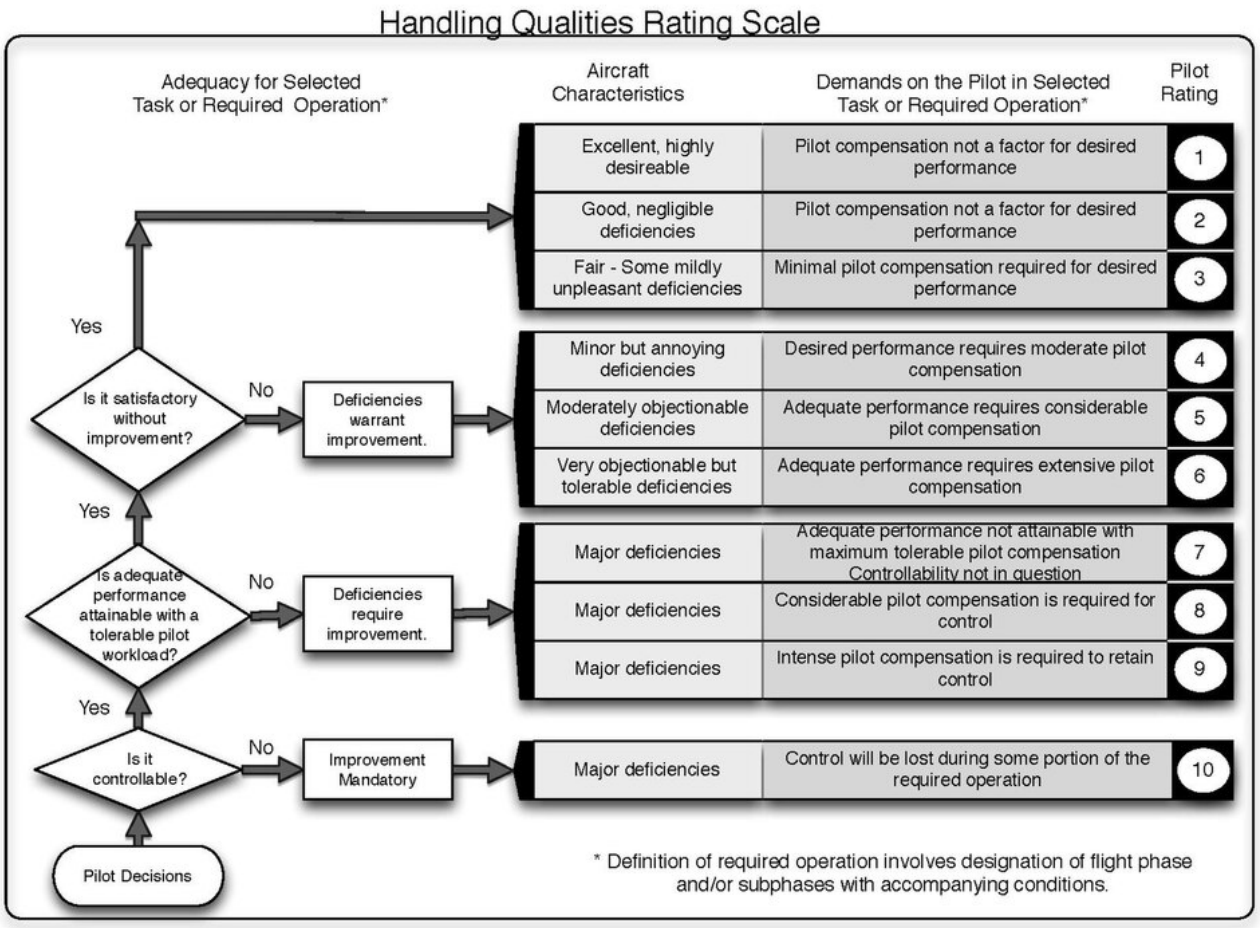
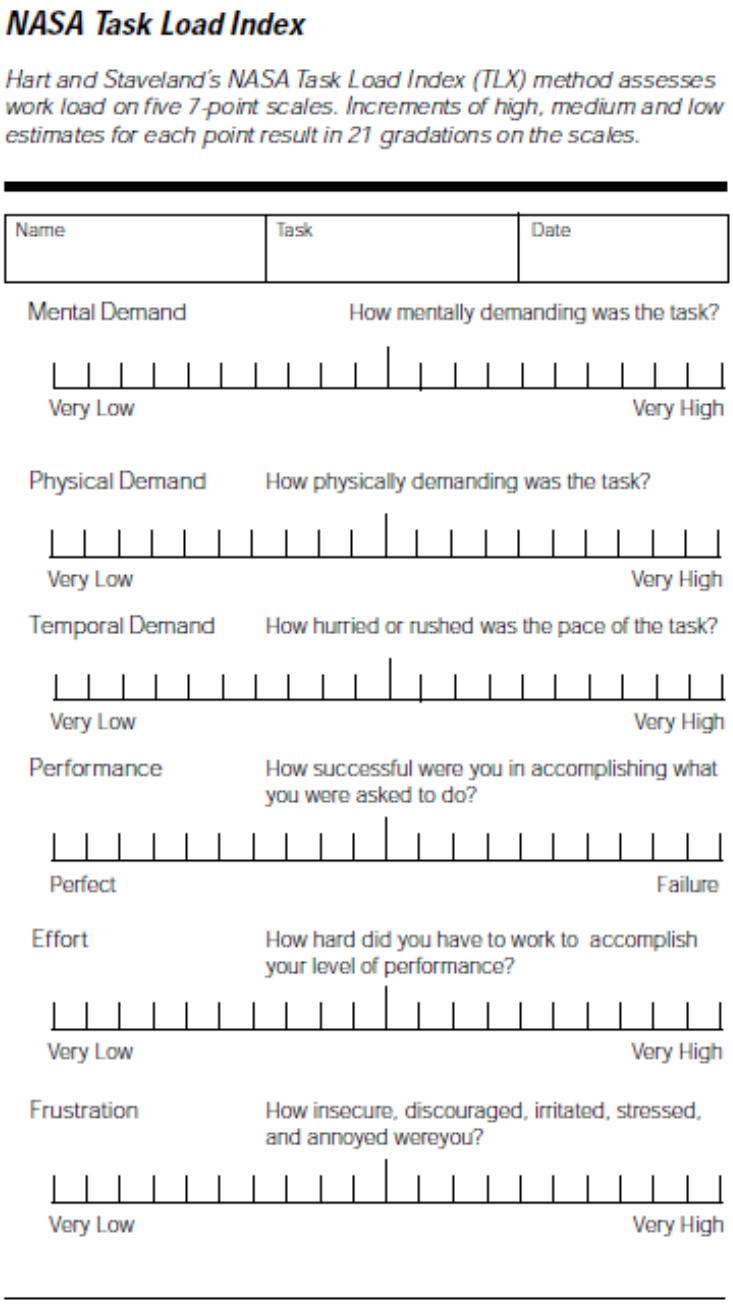


Figure 36 – Pilot setup (flight-test)

C Participant Questionnaire



Handling Qualities Rating Scale



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

NASA Task Load Index