SYSTEMS TECHNOLOGY, INC.
13766 Hawthorne Boulevard, Hawthorne, California 90250-7083, Phone (310) 679-2281, Fax (310) 644-3887

STI Paper No. 784

The Real-Flight Approach to Assess Flight Simulator Force Cueing Fidelity (STI Paper Series)

August 2013

David H. Klyde
Amanda K. Lampton, Ph.D.
P. Chase Schulze
Daniel J. Alvarez
Systems Technology, Inc.

Robin Johnson, Ph.D.
Advanced Brain Monitoring, Inc.

Leah Rowe
Air Force Research Laboratory

Published as AIAA-2013-5161
AIAA Atmospheric Flight Mechanics Conference
August 19-22, 2013, Boston, MA

AIAA Atmospheric Flight Mechanics (AFM) Conference
August 19-22, 2013, Boston, MA
AIAA 2013-5161

The Real-Flight Approach to Assess Flight Simulator Force Cueing Fidelity

David H. Klyde[*]
Amanda K. Lampton, Ph.D.[†]
P. Chase Schulze[‡]
Daniel J. Alvarez§
Systems Technology, Inc., Hawthorne, CA, 90250

Robin Johnson, Ph.D.[**]
Advanced Brain Monitoring, Inc., Carlsbad, CA, 92008

Leah Rowe[††]
Air Force Research Laboratory, Dayton, OH, 45433

For a full slate of reasons that includes economics and aging fleets, the role of simulation in tactical pilot training continues to expand. For decades researchers have been defining ways to assess simulator fidelity that have produced important results and useful metrics. Yet, true measures of positive transfer of training from simulator to flight remain elusive. To address this need, the Real-Flight process is introduced as a means to provide a full spectrum of quantitative and qualitative measures that can provide a means to assess the transfer of training question. This flight-centered approach uses a suite of task performance, pilot-vehicle system, psychophysiological, and pilot opinion measures that offers an assessment pathway. The Real-Flight method is not intended as a "black box" approach where data goes "in" and an assessment comes "out." Instead, it provides the tools by which the trainer can make the required evaluation. In the program described herein, five test pilot participants flew a set of evaluation tasks approved by Air Force subject matter experts and then repeated the flight test sortie in a fixed base simulator using two types of force cueing inceptors. This paper presents the results for a pitch axis tracking task. The findings indicate that the selected metrics were sensitive to the similarities and differences between flight and simulation and to variations in simulator fidelity.

I. Introduction

Deficits in fidelity of the synthetic environment can limit the effectiveness of tactical flight training. In seeking improvements to this process, past research often compared simulators of varying degrees of fidelity to identify deficiencies and recommend improvements. In contrast, the work described herein featured a flight-centered approach (see Figure 1) that provided a direct comparison of representative training scenarios between actual flight and synthetic environments using in-flight simulators and representative ground-based simulators.[1] The resulting database led to both quantitative and qualitative measures that encompass the complete cognitive assessment spectrum including psychophysiological, pilot opinion, task performance, and pilot-vehicle response metrics.

---

[*] VP Research & Engineering Services, AIAA Associate Fellow.
[†] Staff Engineer, Research, AIAA Member
[‡] Staff Engineer, Research, AIAA Member.
[§] Staff Engineer, Research, AIAA Member.
[**] Director of Psychophysiological Research.
[††] Research Psychologist.

This paper was cleared for public release on 10 June 2013, Case Number 88ABW-2013-2734.

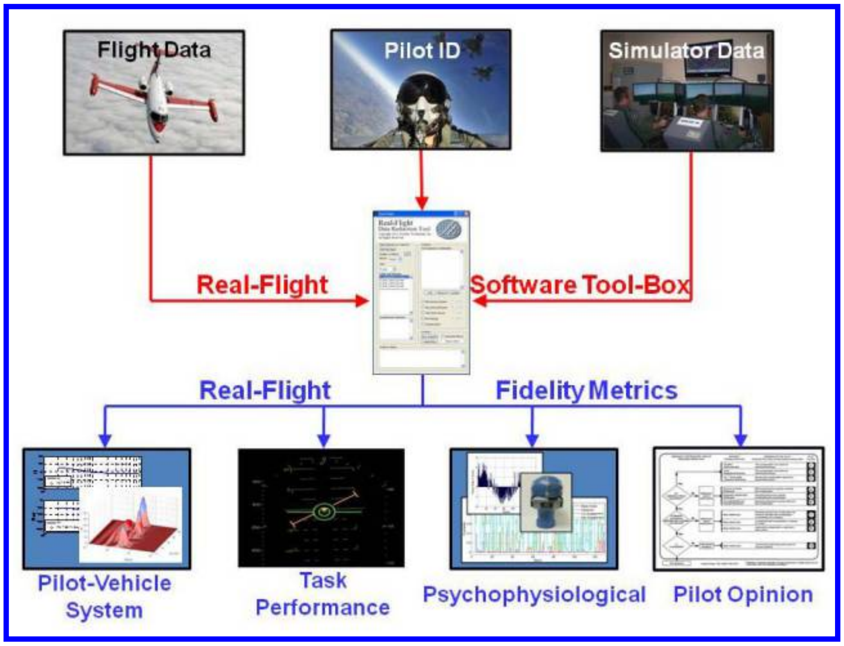


Figure 1: Real-Flight Simulator Fidelity Assessment Tool

Simulators have many components that need to be considered when assessing fidelity for training including the vehicle model, sensor feedback generation, sensory display devices, and the human operator (see Figure 2). The human operator assimilates information from the visual, auditory, proprioceptive, vestibular, and motion cues provided by the simulator sensory display devices (visual displays, speakers, G-seat, motion base, cockpit control inceptors, etc.). To best assess fidelity, significantly pertinent aspects of the human experience when immersed in a simulated world must be considered. The Real-Flight software features a comprehensive set of cognitive and behavioral assessment metrics and host interface that will provide criteria to assess the ability of a given simulator to meet Training and Readiness Requirements. The use of such flight-centered metrics will improve the cost benefit assessment regarding the use of enhanced synthetic environments for training.

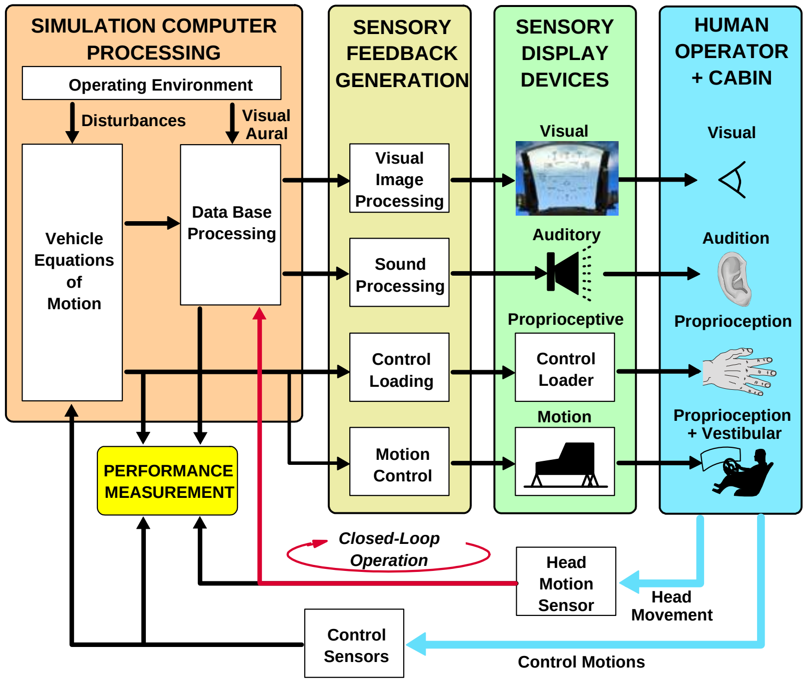


Figure 2: Pilot-in-the-Loop Simulation

The process begins with comparable flight and simulator data from a given pilot participant. The data are comparable in that they are representative of like tasks intended to meet the same training objectives as defined in the appropriate Training and Readiness matrix. The cognitive assessment areas include task performance, pilot-vehicle system response, psychophysiological, and pilot opinion. Specific measures are as follows:

- Task Performance: These are quantitative measures of performance that are defined by the training requirements of a selected task. The team worked with Air Force subject matter experts to map the selected measures to specific Training and Readiness requirements.
- Pilot-Vehicle System Response: Measures are uniquely defined for both continuous and discrete piloting tasks that include time and frequency domain parameters. Examples include crossover frequency, phase margin, peak and average input/output power, rise time, percent overshoot, and settle time.
- Psychophysiological: Measures include EEG (electroencephalogram to measure brain waves) and ECG (electrocardiogram to measure heart rate variability) to assess pilot engagement, distraction, workload, and task speed. Team member Advanced Brain Monitoring, Inc. (ABM) has validated such measures using their B-Alert hardware and data reduction software as described later in this paper.
- Pilot Opinion: Ratings, comments, and debrief questionnaires that address task performance and workload related to specific Training and Readiness Requirements. Because of the expense associated with flight and, to a slightly less extent simulator tests, rarely are enough evaluation runs accumulated from a large enough pool of pilots to generate statistically relevant results. As a result, greater emphasis is placed on the ratings and comments generated by the pilot participants, and thus the methods used to generate such ratings will follow industry best practices as are routinely used by STI and Calspan in related work.

II. Flight Test and Piloted Simulation

A. Aircraft Configurations

Three longitudinal configurations were used in this program. The configurations were defined for a flight condition of 15,000 ft altitude and 250 KIAS. The configurations are identified as follows:

- Lon_2D – Borderline Level 1/2 pitch configuration that is highly PIO resistant even in the presence of significant rate limiting. Initial response may be considered too abrupt to consistently generate Level 1 ratings.
- Lon_2P – Level 2 pitch configuration that has the same bare airframe dynamics as 2D with a command path filter (τ = 0.25). This configuration may have a tendency for undesirable motions when performing high gain tasks.
- Lon_2H – Borderline Level 2/3 pitch configuration that has the same bare airframe dynamics as 2D with a command path filter (τ = 0.5). This configuration is susceptible to PIO when performing high gain tasks.

The key longitudinal control system elements downstream of the feel system are identified in Figure 3. These elements include a command path filter, software rate limiter, surface actuator, surface position limits, and feedback control gains. The command path filter is a first order filter that is used to create two of the longitudinal configurations defined above. The control surface actuator dynamics and surface position limits are those found on the Calspan Learjet. Parameters for the control system elements are defined in Table 1.

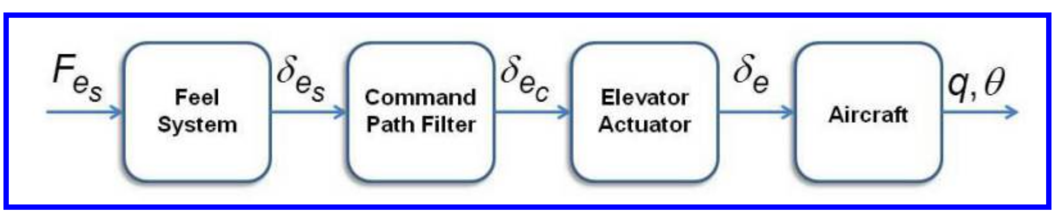


Figure 3: Longitudinal axis control system elements.

Table 1: Longitudinal axis flight control system parameters.

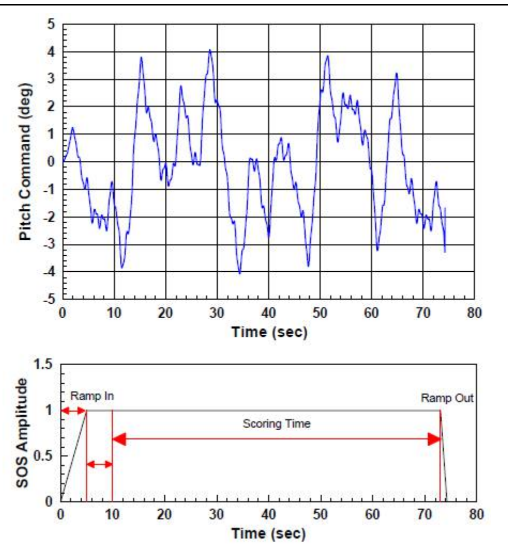
Flight Control System Element	Form	Parameter Values
Feel System	$\frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$	$\omega_n = 20$ r/s; $\zeta = 0.7$
Command Path Filter	$\frac{k_{LON}}{\tau_{LON} s + 1}$	k_{LON} = 1(baseline), varies τ_{LON} = 0 (baseline), 0.25, 0.5
Elevator Actuator	$\frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}$	$\omega_n = 75$ r/s; $\zeta = 0.7$

B. Evaluation Tasks

A set of piloting tasks that are easily repeatable in both the flight test and simulator environments were identified such that direct comparisons between the two environments could be made. The tasks defined herein have been modified from the catalog of fixed wing handling qualities demonstration maneuvers[2] that were assembled as part of a program conducted by STI for the Flight Dynamics Directorate of the Air Force Research Laboratory.[3] The five selected tasks (bank angle capture and hold, pitch attitude capture and hold, pitch and roll attitude sum-of-sines tracking, and discrete pitch and roll attitude tracking) were vetted by subject matter experts from the Human Effectiveness Directorate of the Air Force Research Laboratory prior to their inclusion in the formal test plans that were submitted for Air Force approval. This paper focuses on the results obtained from the pitch axis sum-of-sines tracking task – a surrogate for air-to-air tracking. The sum-of-sines command signal is defined in Table 2 and the associated pilot task description follows below. In the aircraft, the task was flown using a head-down display, while in the simulator a projected head-up display that has been incorporated in the visual display was used.

Table 2: Pitch sum-of-sines command signal.

Sine-wave No.	Pitch Command Signal Parameters		
	A_i (deg)	No. Cycles	ω_i (r/s)
1	-1	2	0.19947
2	1	5	0.49867
3	1	9	0.8976
4	0.5	14	1.39626
5	-0.2	24	2.39359
6	0.2	42	4.18879
7	-0.08	90	8.97597



Task Objectives

- Evaluate handling qualities in a tight, closed-loop tracking task.
- Evaluate feel system and control sensitivity characteristics.
- Identify bobble or PIO tendencies.

Description

Aggressively track the displayed pitch attitude command signal and attempt to keep errors within the specified tolerances.

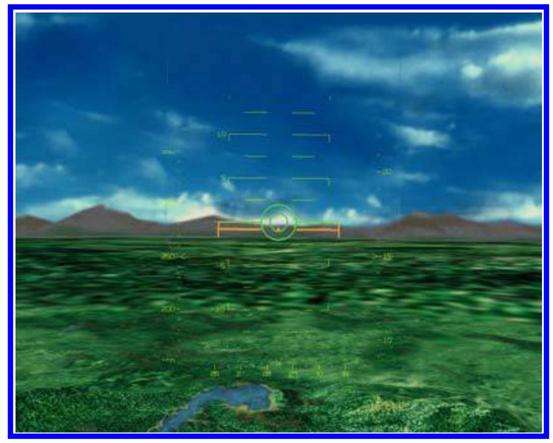
Desired Performance

- ±1° in pitch 50% of the time.
- ±5 kts deviation in airspeed.

Adequate Performance

- ±2° in pitch 50% of the time.
- ±10 kts deviation in airspeed.

a) Learjet head-down display



b) Simulator projected display with

Figure 4: Pitch attitude tracking task display.

C. Flight Test

1. Test Aircraft Description

The flight test sorties were conducted in the Calspan Corporation Learjet 3 variable stability in-flight simulator as shown in Figure 5. The Calspan Learjets provide three degree-of-freedom (3-DOF) in-flight simulation capabilities for advanced stability, control, flying qualities demonstrations, Advanced Maneuvering-Upset Recovery Training (AM-URT), and research. They are also used to test/demonstrate advanced flight control systems concepts. The three aircraft are used in these capacities to support flight test training of test pilots and flight test engineers around the world, as well as to support new aircraft development programs.

Figure 5: Calspan Learjet 3 In-Flight Simulator (photo courtesy of Calspan Corp.).

The right seats of the Learjets have been extensively modified to serve as the Evaluation Pilot (EP) crew station. The normal Learjet wheel/column has been removed. In Lear 3, it is replaced with one of two experimental controllers; 1) centerstick or 2) wheel/column. Each of these two axis control inceptors has programmable variable feel capability, allowing simulation and evaluation of a wide range of characteristics. The Learjet's rudder pedals have also been replaced with variable feel capability. Electrohydraulic servo actuators drive the aircraft's primary control surfaces in response to pilot inputs and the signals from the Variable Stability System (VSS), essentially a programmable fly-by-wire system. The Safety Pilot (SP), whose controllers remain mechanically connected to the Learjet control surfaces via cables, occupies the left seat. The Learjet features a fail-safe limit monitoring system that ensures preset limits are never exceeded. A safety pilot maintains constant capability to recover the aircraft should operational limits be exceeded. Figure 6 shows the modified aircraft configuration for Lear 3 with the incorporated VSS components. This includes a second flight test engineer station that can include a laptop computer pallet directly to the port side of the starboard engineer station shown in the figure. A third flight test engineer can sit in the most rearward seat for takeoff and landing and then move to a jump seat that is located just behind the two pilots for the evaluations. For the flight tests described herein, the B-Alert data collection was conducted from the port engineer station, while the STI test conductor used the jump seat location to better interact with the pilots and monitor the proceedings.

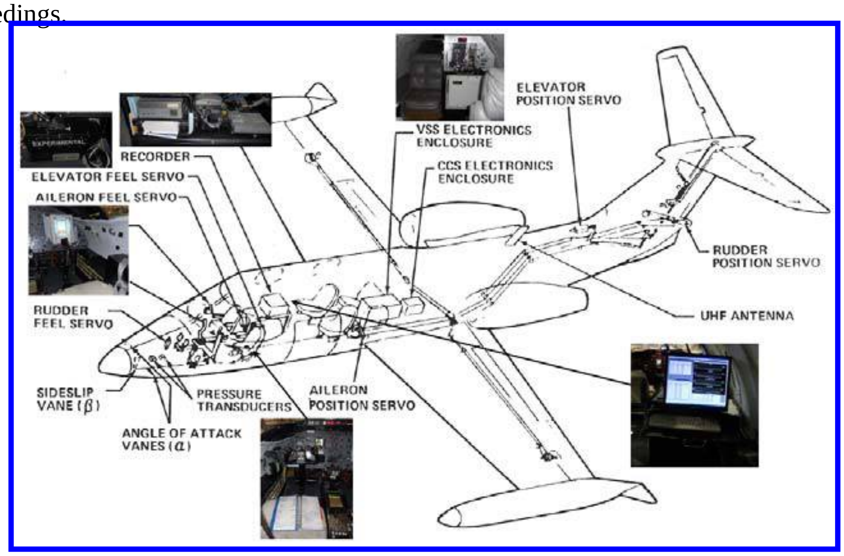


Figure 6: Learjet VSS component layout.

2. Checkout Flight

In preparation for the formal flight test evaluations, a checkout flight was conducted on 3 October 2012 from the Calspan facility at the Niagara Falls, NY airport. The purpose of the checkout flight was to verify a number of critical aspects of the flight test including the vehicle configurations, pilot tasks, and B-Alert EEG/ECG measurement system integration. The vehicle configuration models were verified from in flight data generated from stick force frequency sweeps. The B-Alert system was successfully used in the Learjet and no negative effects were noted either in the aircraft safety-of-flight systems (e.g., electromagnetic interference) or in the quality of the B-Alert data. In addition, the flight test data was successfully synced with the B-Alert data, allowing for the direct comparison of EEG/ECG results with events recorded in the flight data.

3. Evaluation Flights

Five evaluation flights were conducted from the Air Force Test Pilot School facility at Edwards AFB the week on 29 October 2012. On Monday 29 October, four of the five evaluation pilot subjects conducted the required baseline testing for the B-Alert system. The remaining pilot subject performed the test on the Tuesday 30 October. Two evaluation flight sorties were performed on Tuesday 30 October and Wednesday 31 October, while the last flight was performed on Thursday 1 November. For the most part, the test cards were completed for each flight. The number of discrete tracking runs flown in a given sorties was often reduced to accommodate remaining flight time limitations.

D. Piloted Simulation

1. Fixed-Base Simulator Description

The STI fixed base pilot-in-the-loop flight simulator was developed as an additional research tool to strengthen the capabilities of STI in the area of real-time flight simulation and pilot-vehicle system identification. The key elements of the simulator including the pilot are identified in Figure 7. The McFadden feel system is a key component in that it provides flight representative proprioceptive cues that enhance the fidelity of the simulation. To assess the sensitivity of the selected metrics to simulator changes, a lower fidelity fixed spring sidestick controller was also used as part of the evaluation process.

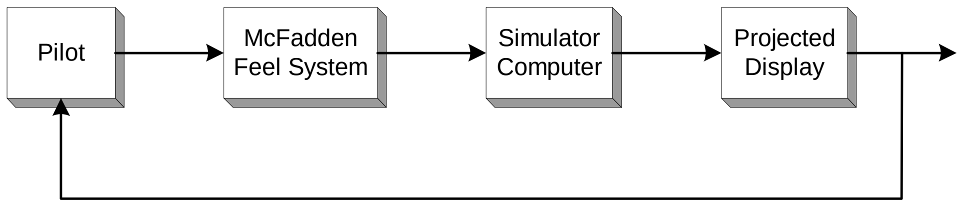


Figure 7: Pilot-in-the-loop simulator elements.

The simulator shown in Figure 8 is a win32 console application designed to interface with Matlab for data input and output. It is capable of simultaneously simulating the time response to arbitrary input of as many linear systems as computer memory will allow. Initialization data for this program is one or more Matlab files, each containing a state-space quadruple representation of a linear dynamic system. Included in this file is information the simulator uses to attach its input and output devices to the input and output states being simulated.

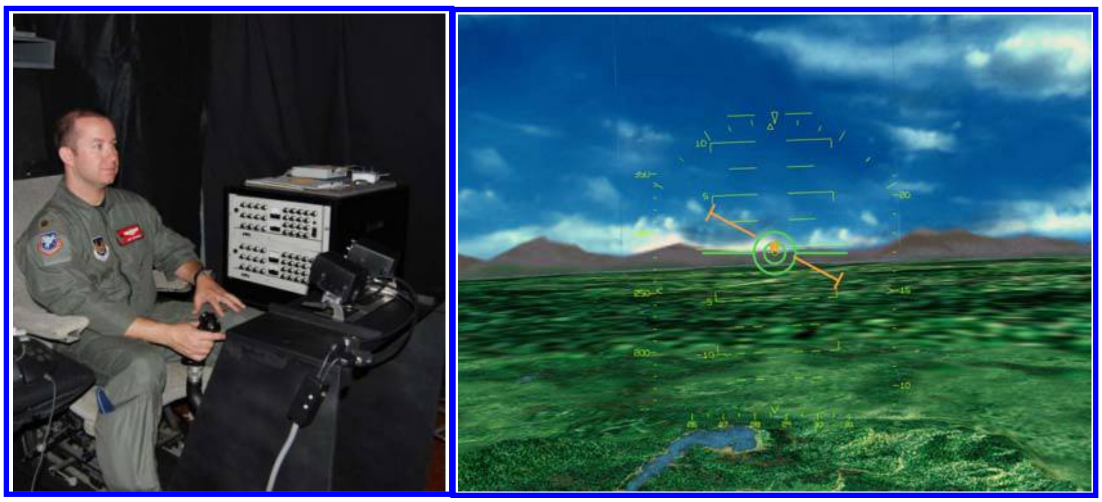


Figure 8: STI simulator with McFadden control loader and projected display.

2. Simulator Evaluations

The simulation test plan was quite simple. First, the complete flight test card for the given pilot was repeated in the simulator with the McFadden control loader. Then a subset of the evaluation tasks, both attitude tracking sum-of-sines tasks, were repeated with the sidestick controller. All piloted simulations were completed within a roughly 2.5 week span following the flight tests. Evaluations for Pilots 3 and 5 were conducted on 5 November 2012, Pilot 2 on 9 November, and Pilots 1 and 4 on 21 November.

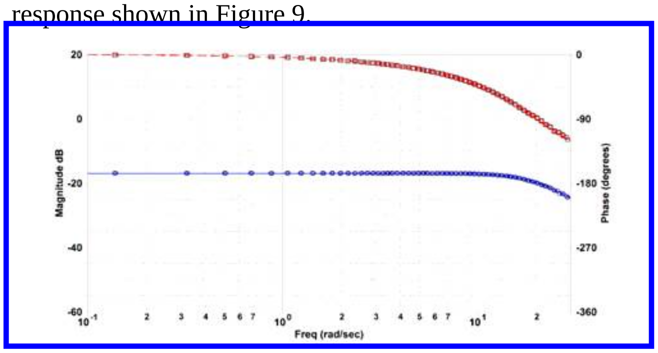
III. Model Comparison

A. Flight Test Configuration Verification

Stick force frequency sweeps (0.1 to 20 r/s) were performed in the Learjet as part of the checkout flight to collect data that was then used to generate frequency responses for the aircraft and feel system dynamics. Transfer function fits to the identified dynamics were used to verify and update the STI fixed based simulator to insure that the models match those used in flight. The identified dynamics are shown in Figure 9 and include the associated fitted transfer functions. These fits closely match the frequency responses for both the feel system and aircraft dynamics. Note that a drop in coherence is responsible for the observed mismatches above 20 r/s in the aircraft responses, which was outside of the range of the frequency sweep input.

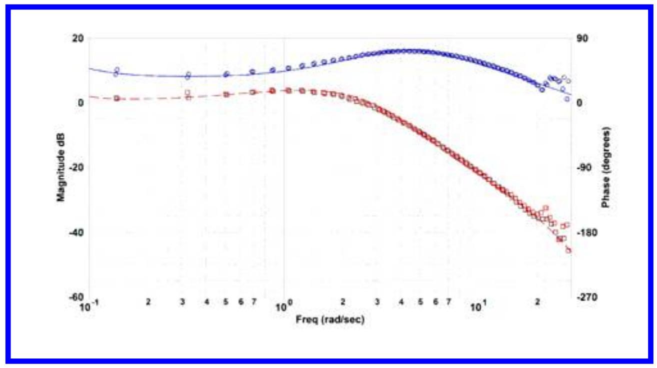
B. Simulator Model Check

A critical element of the flight to simulator comparison was to insure that the aircraft models used in the simulator were representative of those used in flight. Such comparisons are shown in Figure 10 that features the key frequency response, pitch rate to longitudinal stick position. All three longitudinal configurations match extremely well in both magnitude and phase. All that was required to achieve these matches was a command path gain adjustment and proper accounting of the effective time delay that was noted in the flight test identified frequency response shown in Figure 9.



$$\frac{\delta_{es}}{F_{es}} = \frac{57}{[0.7, 20]}$$

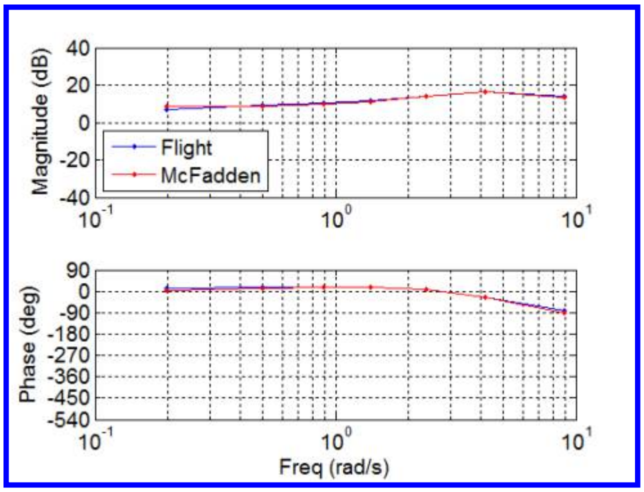
a) Pitch axis feel system



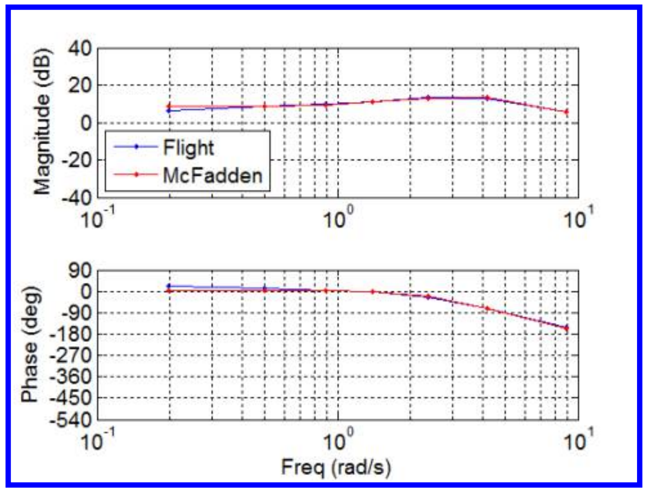
$$\frac{q}{\delta_{es}} = \frac{2.25e5(0)(0.02739)(1.246)}{[0.2389, 0.05637][0.702, 4.571][0.7, 75]} e^{-0.055s}$$

b) Pitch rate to stick position

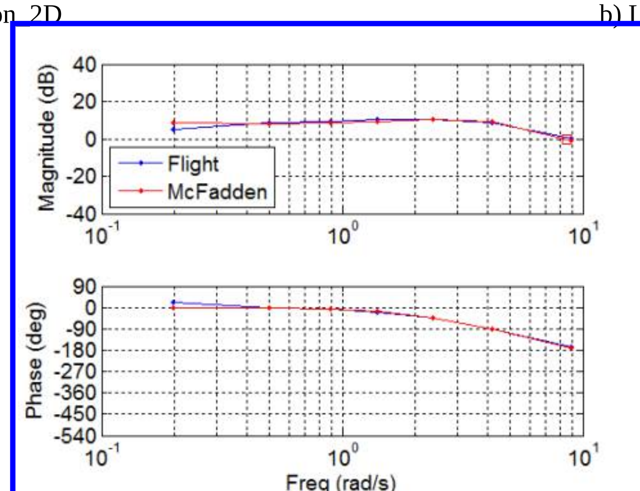
Figure 9: Identified pitch axis dynamics.



a) Lon_2D



b) Lon_2P



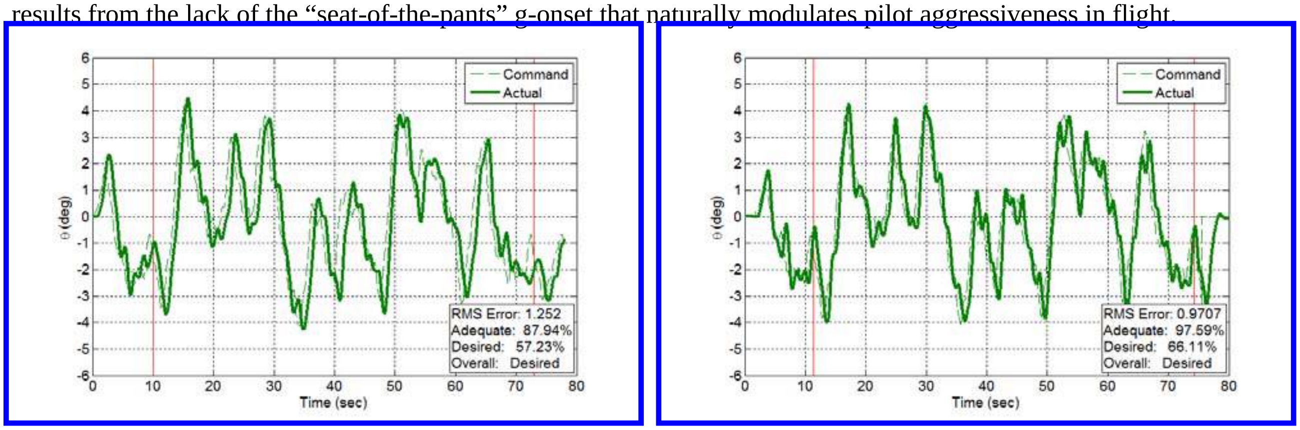
c) Lon_2H

Figure 10: Configuration model comparisons between flight and simulation.

IV. Analytical Results

A. Task Performance

Example pitch axis sum-of-sines tracking performance plots are shown in Figure 11. Here, a side-by-side comparison is shown with flight test results on the left side and simulator results on the right side for Pilot 2 flying the Lon_2H configuration. In the figure, the red vertical lines indicate the task scoring region. Note that in the flight runs, the sine wave phasing was varied run-to-run randomly, but this did not impact the frequency content of the command signal or the resulting task performance comparisons. In the example shown here, the pilot was able to achieve significantly better performance in the fixed base simulator with the McFadden controller. The difference results from the lack of the "seat-of-the-pants" g-onset that naturally modulates pilot aggressiveness in flight.



a) Lon_2H - Flight　　b) Lon_2H – McFadden Controller

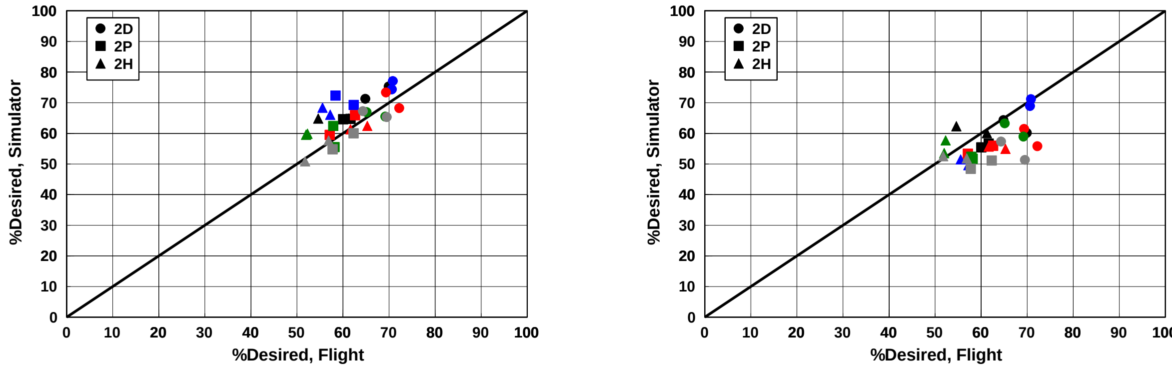
Figure 11: Pitch Attitude Sum-of-Sines Time Histories – Pilot 2

Figure 12 summarizes task performance for all runs by directly comparing simulator versus flight in terms of desired and adequate performance. Desired performance typically fell between 50 and 70%, while adequate performance typically fell above 85%. When comparing the two simulator configurations, McFadden and sidestick, desired performance with the McFadden controller met or exceeded the corresponding flight performance, while for the most part desired performance with the lower fidelity sidestick controller fell below that achieved in flight. There was a little more scatter in the adequate performance results, but the same overall trends were present. The lack of appropriate force cueing from the sidestick controller led the pilots to modify their compensation resulting in degraded performance across all configurations.

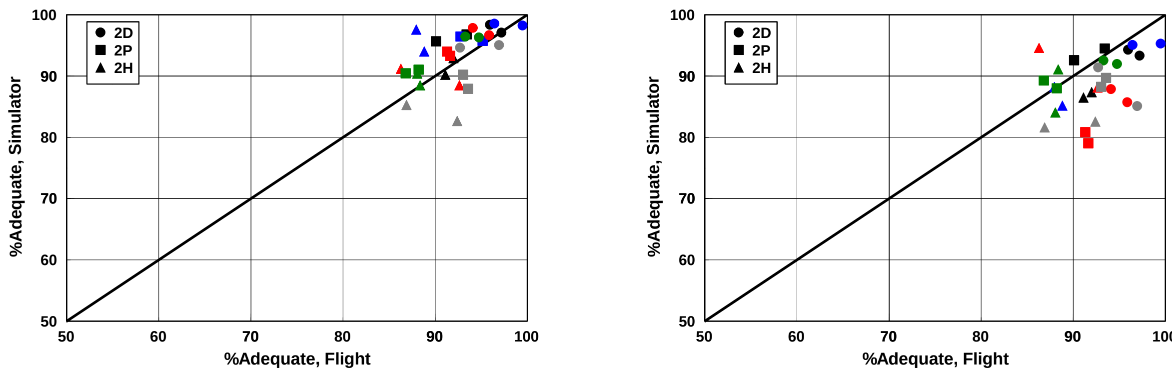
B. Pilot-Vehicle System Measures

For the pitch attitude sum-of-sines tracking cases, the known forcing function input provides a means to directly determine complete pilot-vehicle system and pilot only describing function frequency responses. In this section, the open-loop pilot-vehicle system (i.e., Y_pY_c) and pilot (i.e., Y_p) describing functions obtained in flight are directly compared with those obtained in the simulator using both the McFadden controller and the sidestick. For all cases, the second run conducted for a given configuration was plotted. The first run was considered a familiarity run that allowed the pilot a chance to adopt a compensation strategy, while the second run was considered the primary evaluation run. In the plots, the flight results are shown in blue, while the simulator results are shown in red. Crossover frequencies are indicated on the magnitude responses by a blue circle for flight and a red square for simulator.

A complete set of Y_pY_c and Y_p frequency responses for each pilot were computed and analyzed. Reviewing these plots sheds light into how the pilots adapted to the changing configurations in flight and in the simulator, and more importantly how the pilots adapted to the different force cueing inceptors in the simulator. Classic "k/s-like" behavior as defined by McRuer's Crossover Law[4] was seen in many cases, but was best exemplified by Pilot 5 when flying the Lon_2P configuration (see Figure 13), both in flight and in the simulator.

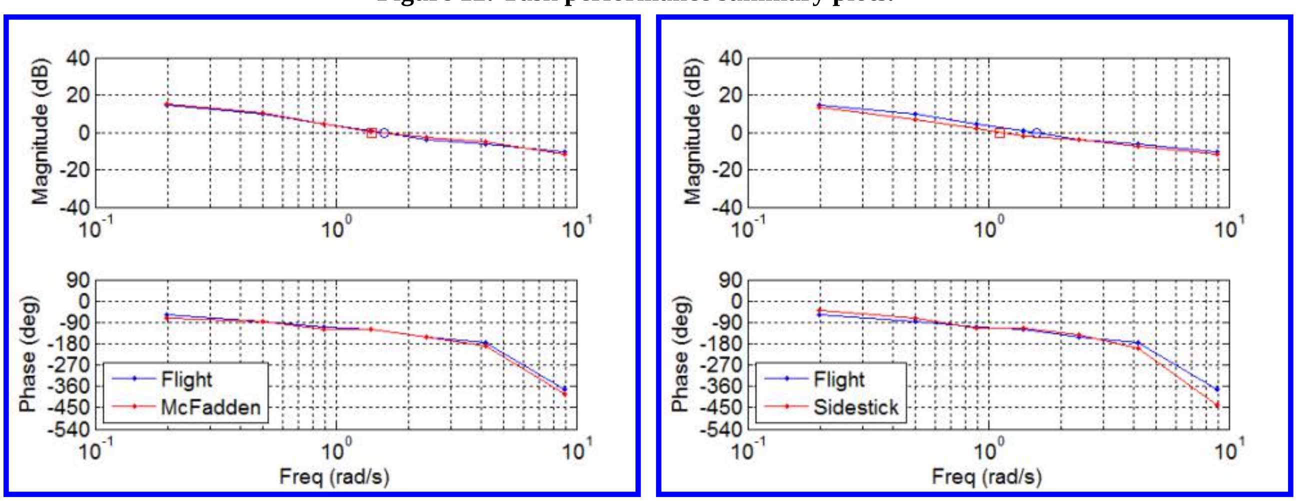


Legend: Pilot 1, Pilot 2, Pilot 3, Pilot 4, Pilot 5



a) McFadden Controller b) Sidestick Controller

Figure 12: Task performance summary plots.

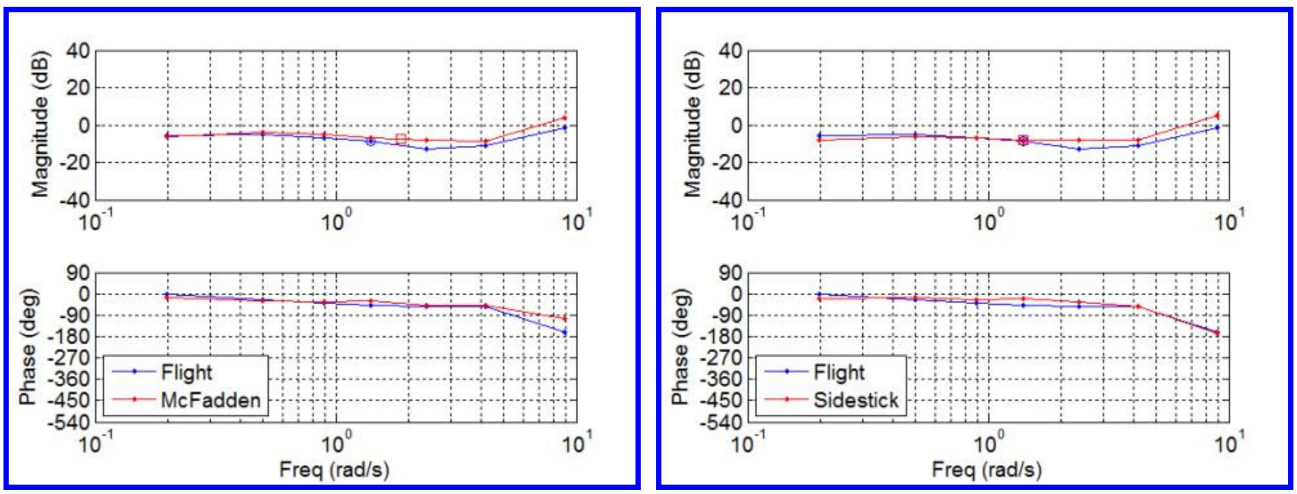


a) Lon_2P – Flight/McFadden b) Lon_2P – Flight/Sidestick

Figure 13: Pilot-vehicle system describing functions for Lon_2P configuration – Pilot 5.

In general, these frequency response comparisons allow, for example, differences in crossover frequencies, pilot gain, pilot lead compensation, etc. to be revealed. For the most part, the overall Y_pY_c and Y_p plots indicated that the pilots maintained similar control strategies between flight and simulator, but differences in pilot gain and lead compensation could be easily observed in the magnitude response plots. As an example consider the Pilot 4 Y_p describing functions for the Lon_2D configuration shown in Figure 14. The red magnitude responses for the two simulator inceptor cases are similar – elevated gain compared to flight and similar lead compensation as indicated by the rise in the magnitude response at higher frequencies. Given these similarities, the pilot was still able to achieve a significantly higher crossover with the higher fidelity McFadden controller.

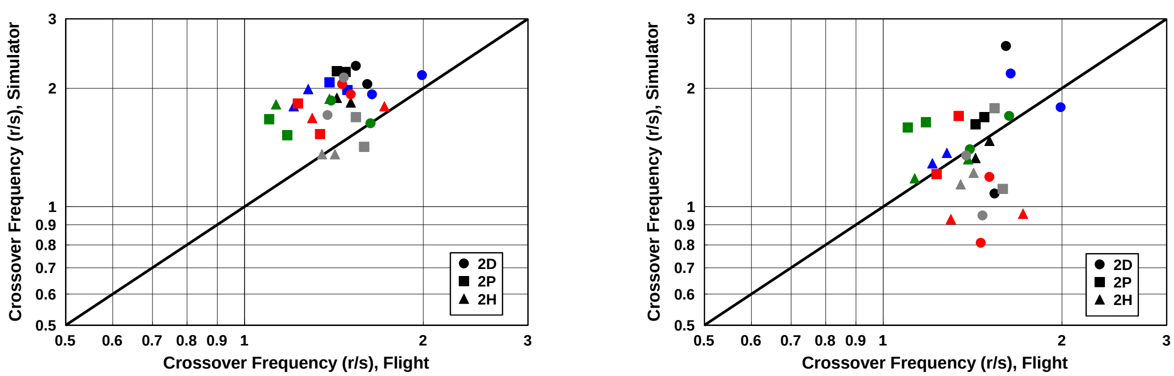
Comparison plots of crossover frequency as obtained in flight versus in the simulator are shown in Figure 15. Here, there are two results for each configuration as the pilot flew the complete task twice in flight and in the simulator. Given an inceptor with reasonable force cueing fidelity, a pilot will typically achieve higher crossover frequencies in a fixed base simulator when compared to flight. This is because the lack of a "seat-of-the-pants" cueing mechanism allows the pilots to fly with a level of aggressiveness not matched in flight. This was indeed the case for the McFadden controller. Much different results, however, were seen with the lower fidelity sidestick where no pilot saw higher crossover frequencies in the simulator for all configurations and runs as was typical with the McFadden controller.



a) Lon_2D – Flight/McFadden

b) Lon_2D – Flight/Sidestick

Figure 14: Pilot describing functions for Lon_2D configuration – Pilot 4.



Legend: Pilot 1, Pilot 2, Pilot 3, Pilot 4, Pilot 5

a) McFadden

b) Sidestick

Figure 15: Pilot-vehicle system crossover frequency summary.

V. Psychophysiological Results

A. Background

There is a distinct human physiological response to stimuli. With proper sensors, this physiological response can be measured, and from these measurements inter- and intra-pilot differences between full motion flight and limited or no motion ground-based simulation environments can be quantified. In this program, the wireless B-Alert Electroencephalogram (EEG) Sensor Headset developed by ABM was used. The B-Alert combines battery-powered hardware with a sensor placement system to provide a lightweight, easy-to-apply method to acquire and analyze up to 24 channels of high-quality EEG plus additional physiological signals such as Electrocardiogram (ECG), respiration, or integrated eye tracking (see Figure 16 for a 10 channel solution that combines 9 EEG channels plus ECG). The sensors require no scalp preparation and provide a comfortable and secure sensor-scalp interface for 8 to 12 hours of continuous use with rechargeable lithium batteries. The headset was designed with fixed sensor locations for three head sizes (e.g., small, medium and large) with placement determined according to the International 10-20 system coordinates. Example sensor site scalp locations on the current x10 system include the following: F3, F4, C3, C4, P3, P4, Fz, Cz, POz (see Figure 16). These sites can be combined in bi-polar or monopolar configurations (referenced to linked mastoids). Amplification, digitization, and radio frequency (RF) transmission of the signals are accomplished with miniaturized electronics in a portable unit worn on the head. The combination of amplification and digitization of the EEG close to the sensors and wireless transmission of the data facilitates the acquisition of high quality signals even in high electromagnetic interference (EMI) environments. Per testing conducted by Calspan Corporation, there were no EMI issues surrounding the use of the wireless B-Alert system in the Learjet.

Figure 16: ABM B-Alert wireless EEG headset.

In addition to high quality EEG, the x10 system includes integrated ECG, allowing assessment of heart-rate variability (HRV) metrics. HRV is a direct measure of autonomic activation, with the low frequency: high frequency ratio providing a metric of anxiety and stress. For example, multiple studies have found a relationship between effort at work, stress, and HRV.[5,6,7] Changes in HRV have also been found to precede changes in stress related cortisol.[8,9]

Psychophysiological measures result from physiological measures that are used to define mental workload metrics. ABM has implemented software for acquisition and real-time analysis of the EEG and ECG data and demonstrated feasibility of operational monitoring of EEG indices of mental workload.[10] Signal analysis techniques were developed to identify and decontaminate fast and slow eye blinks and identify and reject EEG data points contaminated with electromyography (EMG) signals, amplifier saturation, and/or excursions due to movement artifacts. The B-Alert x10 system has the ability to assess workload and distraction over time with built in metrics, such as Engagement, Distraction, Workload, and Task Speed.[11,12] These metrics are exemplified in Figure 17.

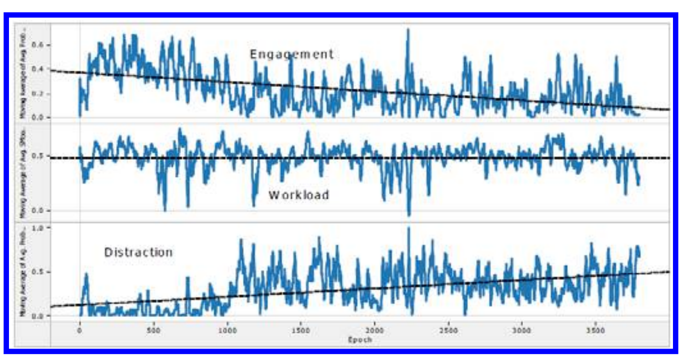


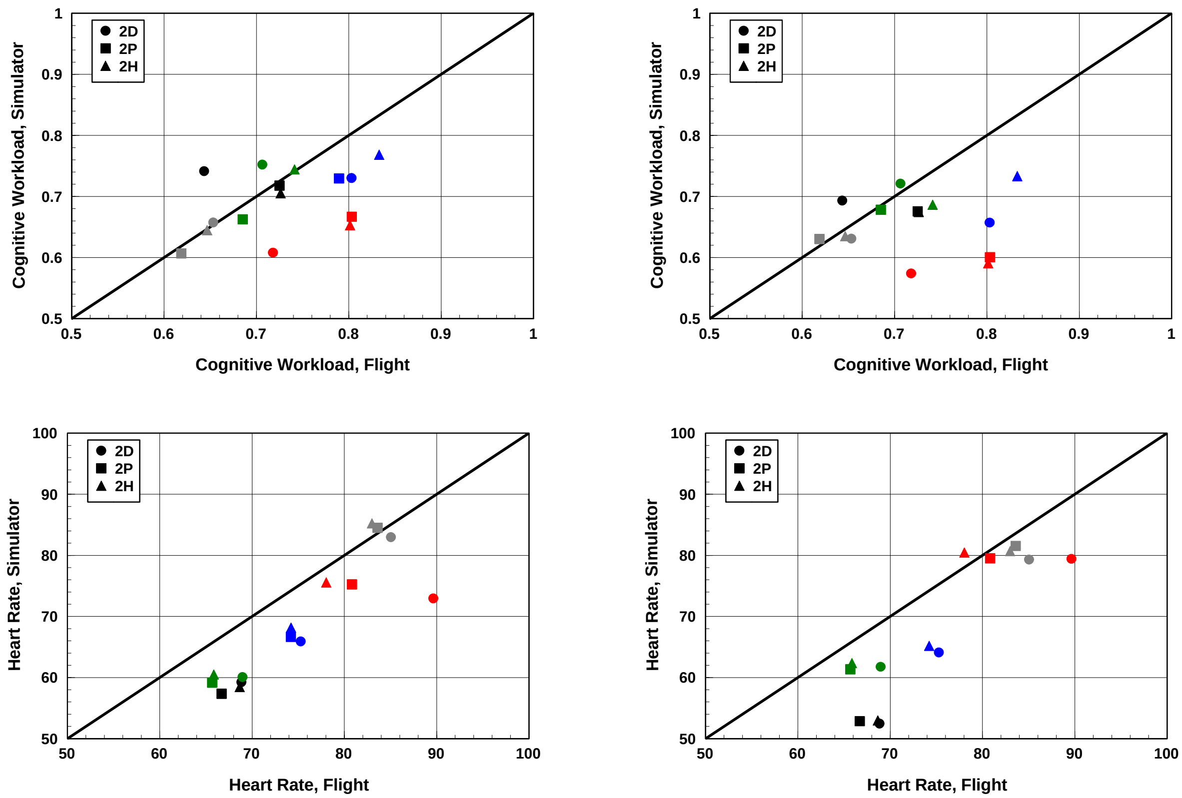
Figure 17: Dynamics of engagement, workload, and distraction over a training session.

B. Individual Pilot Results

There were no issues, complications, or problems associated with the collection of psychophysiological data using the ABM B-Alert headset in flight or in the simulator. The wireless Bluetooth system worked flawlessly in both environments. There were no EMI issues in the flight environment. The B-Alert headset did not interfere with the communication headset used in the Learjet with regard to function and comfort. Relatively simple data handling methods allowed for the EEG/ECG data to be synced with the aircraft flight test data. One surprise of note when this data was examined was that there was a distinct linear offset in heart rate for all tasks. The effect was stronger for some cases (see Figure 18a), but the observation held across the tasks. If the trend held for a larger sample size, it would be possible to predict the change in heart rate between flight and simulator for a selected pilot. The workload results, however, were more varied.

The specific Pitch SOS tracking results are shown in Figure 18 for both the McFadden and sidestick controllers. First, it is observed that the measured workload for Pilot 3 was significantly higher in flight. This trend was also noted for Pilot 2, but to a lesser extent than Pilot 3. The differences observed between flight and simulator for these two pilots was greater with the sidestick controller. For the other three pilots, the workload results compared well between flight and simulator.

A Z-Score, or standard score, indicates how many standard deviations a data point is above or below the mean. This dimensionless quantity is determined by subtracting the population mean from the individual's raw score and then dividing the result by the population standard deviation. For the B-Alert data, the raw pilot scores are standardized by the baseline test results. Specifically, the 3CVT, a moderate workload task, mean and standard deviation were applied to the flight and simulation tasks to address issues associated with range variability across individuals. The workload algorithm range can vary greatly across individuals, with one individual having a normal range from, for example, 0.1 to 0.9, while another might only vary from 0.4 to 0.7. To equalize the "low" and "high" end, z-scoring was used to adjust each pilot's data to his own range, in units of standard deviations from his norm, thus making more equal comparisons across individuals possible.

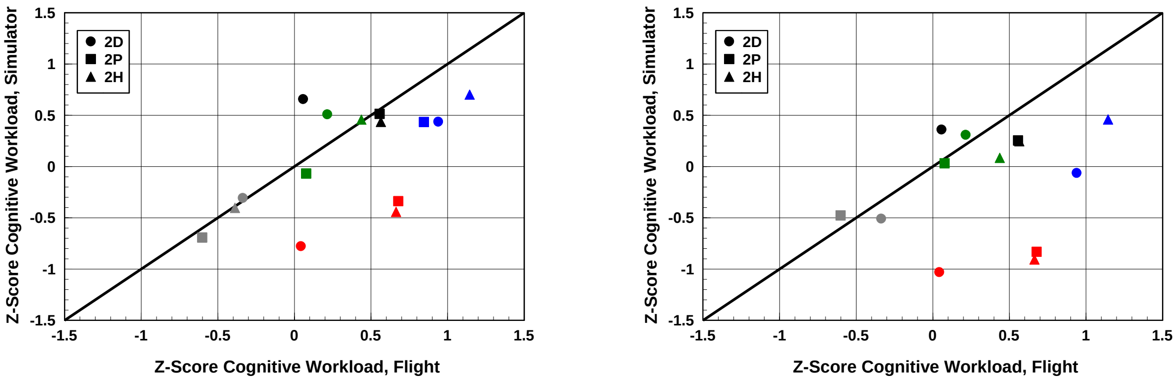


Legend: Pilot 1, Pilot 2, Pilot 3, Pilot 4, Pilot 5

a) McFadden b) Sidestick

Figure 18: Cognitive workload and heart rate for pitch sum-of-sines tracking task.

The Z-Scored cognitive workload results for the pitch sum-of-sines tracking tasks are shown in Figure 21. Here again, both the McFadden controller and the sidestick results are compared to flight. As seen in the above analysis, for all cases the Pilot 3 results showed significantly higher workload in flight compared to either simulator condition. For the other four pilots, however, the McFadden controller simulator condition matched the flight results well. The results were not as good with the sidestick controller. Even Pilot 3 diverged further from the flight results with this inceptor, so clearly the fidelity of the force cueing device was important and can be measured.



Legend: Pilot 1, Pilot 2, Pilot 3, Pilot 4, Pilot 5

a) McFadden b) Sidestick

Figure 19: Z-Scored cognitive workload for pitch sum-of-sines tracking task.

C. Analysis of Variations

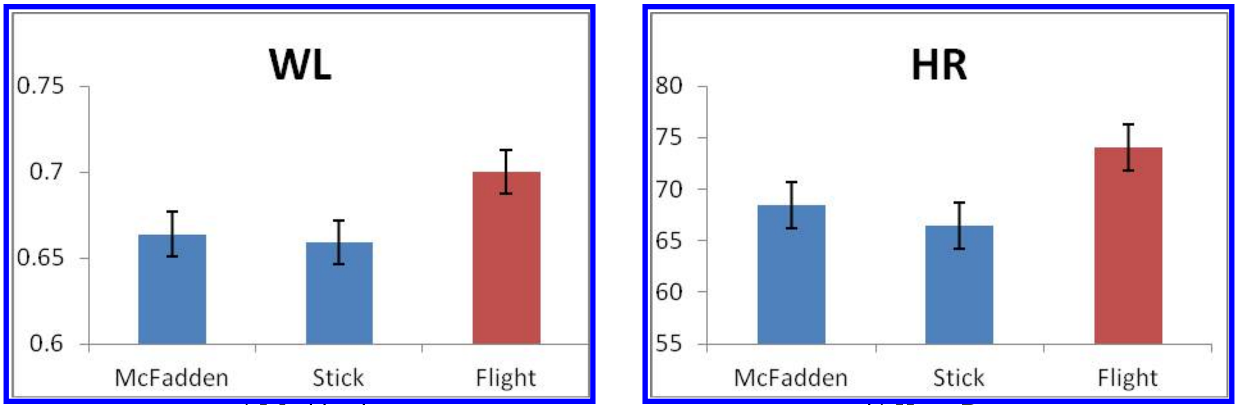
An analysis of variance (ANOVA) was conducted to examine the main effects of aircraft configuration, task (BACH, PACH PSoS, RSoS, and DT,)*, and condition – the pilots were in the fixed base simulator with either the McFadden controller or Sidestick or in flight. Thus, the ANOVA featured 5 configurations, 5 tasks, 3 conditions and a total of 169 observations. Two way and three way interactions were also examined. Dependant variables included heart rate, B-Alert classifiers (i.e., Drowsiness, Engagement, Workload), and power spectra density (PSD)-based bandwidths (i.e., Theta, Alpha, Beta, Gamma) at each of the 9 EEG channel scalp sites (Fz, Cz, POz, F3, F4, C3, C4, P3, P4) as identified in Figure 16. The ANOVA results are summarized as follows:

- Flight leads to an overall increase in physiological activation compared to both simulation conditions including an overall increase in EEG power across all bandwidths and scalp sites.
- Flight leads to a significant, but small increase in EEG-workload.
- Power spectral elevations across task and configuration can be primarily explained by the main effects (aircraft configuration, task, and condition) of overall elevations in the PSDs.

Taken as a whole these data indicate that while there are main effects for flight being associated with overall elevation in activation and workload, in fact there are a great many interactions across the configurations and tasks that indicate that the in flight experience is not substantially different than the simulation experiences.

No significant difference in the B-Alert metric of engagement was shown. However, flight was associated with greater workload and elevated heart rate as shown in Figure 20. Included on the bars is a representation of the standard error of the mean, i.e., the standard deviation of the error in the sample mean relative to the true mean. It should be noted that these differences, while statistically significant are not substantial, with the increase in workload at 0.05 in probability, and the increase in heart rate less than 6 bpm. In addition, there is an overall increase in EEG power across all scalp sites, in all bandwidths of approximately 0.2 μV^2. Broadly these changes are also relatively minor, although consistent across pilots in this study. These data indicate that flight may increase overall physiological activation, and result in slightly elevated perception of workload. It is, however, not substantively different than flight simulation.

* BACH – Bank Angle Capture & Hold, PACH – Pitch Attitude Capture and Hold, PSoS – Pitch Sum-of-Sines, RSoS – Roll Sum-of-Sines, and DT – Discrete Tracking



a) Workload b) Heart Rate

Figure 20: Workload and Heart Rate as a Function of Condition

VI. Pilot Opinion Results

A. Ratings

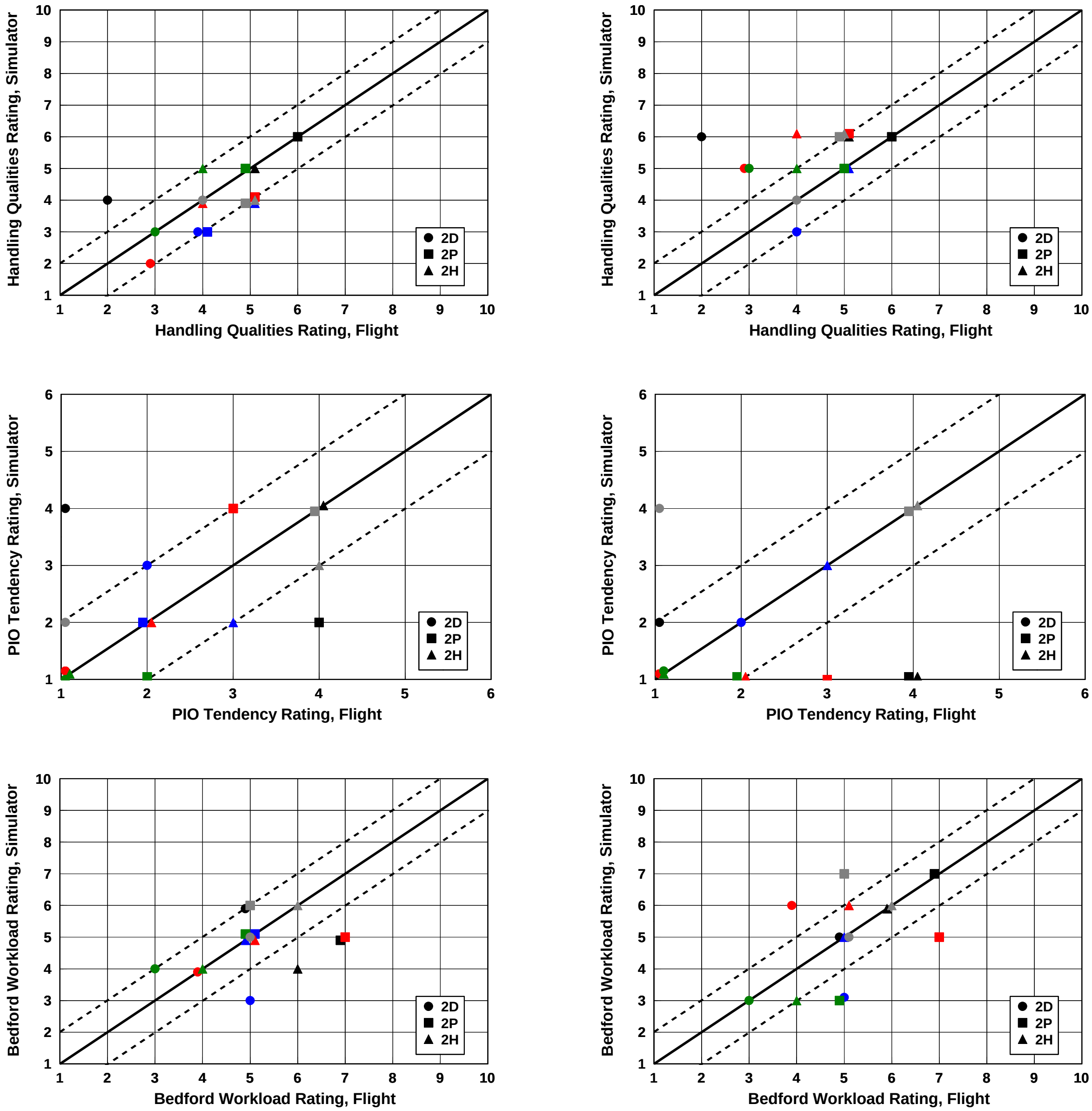
Three pilot rating scales were used in this program. Two were intimately familiar to the test pilot participants, the Cooper-Harper Handling Qualities Rating Scale and the Pilot-Induced Oscillation (PIO) Tendency Scale, while the third, the Bedford Workload Scale, was new to most. The Cooper-Harper scale[13] has been used successfully for both piloted simulation and flight test evaluations for over 40 years. Why has this scale been so successful? First, it accounts for all elements of the control loop that can influence handling qualities; task description, pilot, cockpit interface, stability and control characteristics (i.e., aircraft dynamics), aircraft environment, and task performance. Furthermore, the scale employs an ingenious decision tree that guides the pilot through the scale to the appropriate rating. Finally, it directly connects aircraft characteristics with required pilot compensation (i.e., a measure of pilot workload).

In the same time period that various handling qualities rating scales including Cooper-Harper were under development, PIO tendency rating scales were also evolving. A version of the evolving 6-point scale created by Cornell Aeronautical Laboratory was incorporated into the supporting documentation for the MIL-F-8785B release of the military flying qualities specification.[14] In the early 1980's, Calspan Corporation introduced a revised 6-point scale that introduced a helpful decision tree format[15], but unfortunately removed much of the important descriptive material from the military standard version. To alleviate this issue, a hybrid version of the scale has been created that not only retains the decision tree format but also restores the important descriptions. This version of the scale has been used successfully in recent flight test programs.[16,17]

A gold standard for flight and simulation testing is to find a means to quantify pilot workload. Some 20 years after Cooper-Harper, Roscoe and Ellis created the Bedford Workload Rating Scale as a means to qualitatively address this need.[18] The resulting 10-point scale, derived from the Cooper-Harper scale, features a decision tree that guides the pilot to a workload rating. The Bedford Scale addresses the excess capacity of a pilot to handle secondary tasks. It also leaves room for the pilot to consider physical rather than just cognitive workload as part of the overall rating.

The pitch attitude sum-of-sines tracking pilot rating results are shown in Figure 21. Each pilot rated the three configurations in both flight and the simulator thus resulting in 15 total rating points. The figures feature not only a one-to-one (solid) line, but also plus and minus one rating point (dashed) lines. In an ideal case, the rating comparisons would all fall along the one-to-one line. Practically speaking, rating comparisons that fell within the one rating point threshold were considered "good" comparisons. For the cases with the McFadden controller, both HQR (14/15 points) and PIO tendency ratings (13/15 points) compare well with flight. Only ratings from Pilot 1 fall out of the one rating point threshold. The same is not true for the sidestick where the pilots had to significantly alter their technique to account for the less favorable inceptor characteristics.

Workload ratings generally followed similar trends, but with a greater number falling outside of the threshold. This also reflects a greater difficulty in interpreting this scale when compared to the more familiar Cooper-Harper and PIO Tendency scales. It is interesting that for the McFadden cases, those outside the range were only lower in the simulator when compared to flight, but also note that more than half of the cases (8/15 points) had a one to one comparison with flight. There was more scatter above and below the one rating point threshold with the sidestick, but the workload results were not markedly worse than those obtained with the McFadden controller.



Legend: Pilot 1, Pilot 2, Pilot 3, Pilot 4, Pilot 5

a) McFadden Controller

b) Sidestick Controller

Figure 21: Pitch Attitude Sum-of-Sines Tracking Pilot Ratings

B. Debrief Questionnaire

At the conclusion of each piloted simulation session, the test pilot participant responded to the debrief questionnaire shown in Figure 22. The questionnaire featured a set of four questions concerning the McFadden controller, sidestick, projected display, and B-Alert headset. The results are shown in Figure 23. First, all five pilots agreed or strongly agreed that the McFadden controller provided force cues, workload, compensation, and performance that was representative of flight. The same was not true for the sidestick thus indicating that the quality of the inceptor can significantly influence pilot attitude toward the simulator experience. The responses to the visual display questions were more mixed. Overall the pilots were neutral in this area, so clearly improvements can be made. Such improvements typically include higher resolution graphics and wider field-of-view displays. The final question addressed the use of the B-Alert headset. Four pilots strongly agreed and one agreed that the headset did not interfere with task performance. Thus, the debrief data does expose differences in the force cueing fidelity of the two cockpit inceptors used in the simulator. The takeaway here is that questionnaires remain an important tool for extracting qualitative data from test participants.

McFadden Controller

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
Provided flight representative control force cues;	○	○	○	○	○
Workload was representative of flight;	○	○	○	○	○
Pilot compensation was representative of flight;	○	○	○	○	○
Performance was representative of flight.	○	○	○	○	○

Sidestick

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
Provided flight representative control force cues;	○	○	○	○	○
Workload was representative of flight;	○	○	○	○	○
Pilot compensation was representative of flight;	○	○	○	○	○
Performance was representative of flight.	○	○	○	○	○

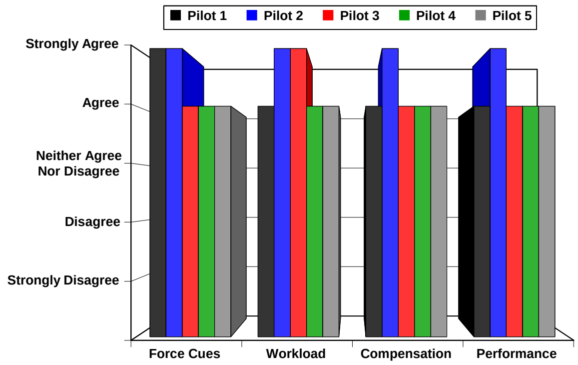
Projected Display

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
Provided flight representative control visual cues;	○	○	○	○	○
Workload was representative of flight;	○	○	○	○	○
Pilot compensation was representative of flight;	○	○	○	○	○
Performance was representative of flight.	○	○	○	○	○

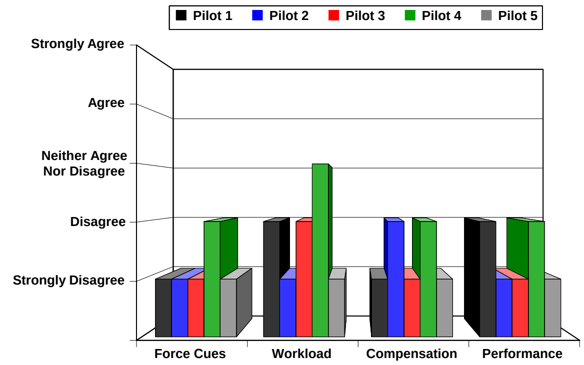
B-Alert Headset

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
Did NOT interfere with ability to perform task.	○	○	○	○	○

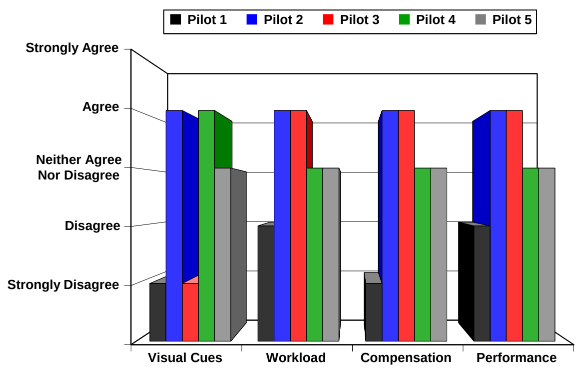
Figure 22: Pilot questionnaire from fixed-base simulator evaluations.



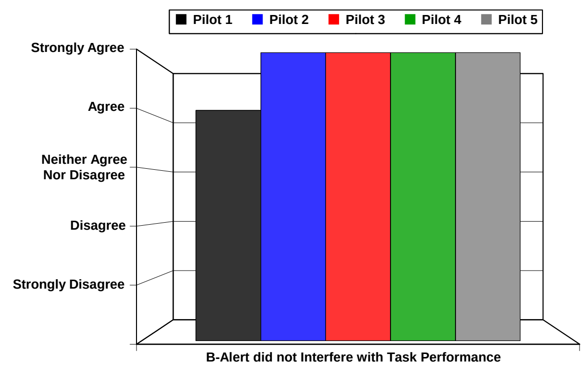
a) McFadden Controller



b) Sidestick Controller



c) Visual Display



d) B-Alert Headset

Figure 23: Summary of pilot questionnaire results.

VII. Conclusions

The purpose of this program was to validate the Real-Flight process as a means to discriminate between flight and simulation. The results of this program were used to address this purpose, not to make a definitive statement regarding the use of simulation for flight training. The authors do feel, however, that the process described herein has now been established through which these questions can be more effectively addressed.

Specifically, in this program cockpit inceptors with different force cueing characteristics were used to assess the Real-Flight metrics. The results indicated that the selected task performance, pilot-vehicle system, psychophysiological, and pilot opinion measures were sensitive to the similarities and differences between flight and simulation and to variations in simulator fidelity. Furthermore, for the first time, EEG/ECG measures of a pilot in command were made in flight and directly compared to those obtained in a simulator with the same pilot performing the same task. The results indicate that for the selected tasks, psychophysiological measures including cognitive workload were elevated in flight, but not by a significant margin. In terms of human opinion, use of pilot ratings with well-established scales and a pool of pilots that have been trained to use them provided a discriminating method to quantify that opinion between flight and simulation.

Regarding the Real-Flight software toolbox, it is not intended to be a “black box” solution where data goes “in” and a pass/fail judgment comes “out.” Instead it provides a comprehensive set of measures that can be used to assess how well a given simulator setup compares to flight. In the end, how well the simulator performs will be open to interpretation of the data, but the goal of the Real-Flight process is to provide the processed data with guidelines to interpretation so that opinion alone does not dominate the assessment. The Real-Flight process can also be used to compare data between simulator configurations not just between simulator and flight. The examples documented herein clearly exposed the similarities and differences between the McFadden controller and the sidestick, clearly highlighting the benefits of the higher fidelity McFadden inceptor.

Acknowledgements

The authors of this paper would first like to acknowledge the contributions of 2dLt Robert Nelson, 2dLt Thomas McNitt, Capt Danny Lacore, Gary Hellman, Dr. Sharon Conwell, and Kim London of the Air Force Research Laboratory without whom the required approvals would still be pending. Next, William Gray and Nathan Cook were instrumental in coordinating the participation of the Air Force Test Pilot School. Finally, the work could not have been completed without the exceptional efforts of Ryan McMahon, Jason Kirkpatrick, Evan Thomas, Kevin Prosser, Paul Deppe, and Paul Schifferle of Calspan Corporation.

References

1. Klyde, D. H., J. G. Howe, R. W. Allen, M. McCauley, P. M. Crane, "A Flight Centered Approach to Assess Dynamic Flight Simulator Force Cueing," AIAA 2011-921 presented at the *49th Aerospace Sciences Meeting*, Orlando, FL, 4-7 Jan. 2011.
2. Klyde, D. H., and D. G. Mitchell, *Handling Qualities Demonstration Maneuvers for Fixed Wing Aircraft Vol. II: Maneuver Catalog*, WL-TR-97-3100, Oct. 1997.
3. Klyde, D. H., B. L. Aponso, and D. G. Mitchell, *Handling Qualities Demonstration Maneuvers for Fixed-Wing Aircraft Volume I: Maneuver Development Process*, WL-TR-97-3099, Oct. 1997.
4. McRuer, D. T., and E. S. Krendel, *Mathematical Models of Human Pilot Behavior*, AGARDograph No. 188, Nov. 1973.
5. Uusitalo, A., T. Mets, K. Martinmaki, S. Mauno, U. Kinnunen, and H. Rusko, "Heart Rate Variability Related to Effort at Work," *Appl Ergon*, 2011.
6. Karasek, R., S. Collins, E. Clays, A. Bortkiewicz, and M. Ferrario, "Description of a Large-Scale Study Design to Assess Work-Stress-Disease Associations for Cardiovascular Disease," *International Journal Occupational Medicine and Environmental Health*, Vol. 23, No. 3, 2010, pp. 293-312.
7. Lo, S.H., L.Y. Lin, J.S. Hwang, Y.Y. Chang, C.S. Liau, and J.D. Wang, "Working the Night Shift Causes Increased Vascular Stress and Delayed Recovery in Young Women," *Chronobiol Int.*, Vol. 27, No. 7, 2010, pp. 1454-68.
8. Hynynen, E., N. Konttinen, and H. Rusko, "Heart Rate Variability and Stress Hormones in Novice and Experienced Parachutists Anticipating a Jump," *Aviation Space Environmental Medicine*, Vol. 80, No. 11, 2009, pp. 976-80.
9. Ledowski, T., B. Bein, R. Hanss, A. Paris, W. Fudickar, J. Scholz, et al., "Neuroendocrine Stress Response and Heart Rate Variability: a Comparison of Total Intravenous versus Balanced Anesthesia," *Anesth Analg*, Vol. 101, No. 6, 2005, pp. 1700-5.
10. Berka, C., D. Levendowski, C.K. Ramsey, et al., "Evaluation of an EEG-Workload Model in an Aegis Simulation Environment," *Proceedings: Biomonitoring for Physiological & Cognitive Performance during Military Operations*, J.A. Caldwell and N.J. Wesensten, editors, Orlando, FL, 2005, pp. 90-99.
11. Berka, C., D. Levendowski, M. Lumincao, A. Yau, G. Davis, and V. Zivkovic, "EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning and Memory Tasks," *Aviat Space Environ Med*, Vol. 78, No. 5, 2007, pp. B231-B44.
12. Johnson, R. R., D. Popovic, R. E. Olmstead, M. Stikic, D.J. Levendowski, and C. Berka, "Drowsiness Determination through EEG: Development and Validation," *Biol Psychol.*, Vol. 87, No. 2, 2011, pp. 241-250.
13. Cooper, George E., and Robert P. Harper, Jr., *The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities*, NASA TN D-5153, April 1969.
14. Chalk, C. R., T. P. Neal, T. M. Harris, F. E. Pritchard, Background Information and User Guide for MIL-F-8785B (ASG), "Military Specification – Flying Qualities of Piloted Airplanes," AFFDL-TR-69-72, Aug. 1969.
15. Weingarten, N. C., and C. R. Chalk, In-Flight Investigation of Large Airplane Flying Qualities for Approach and Landing, AFWAL-TR-81-3118, September 1981.
16. Klyde, D. H., J. G. Reinsberg, E. Sanders, and A. Kokolios, "Flight Test Evaluation of Stability Augmentation Steering System for Aircraft Ground Handling," *J. Guidance, Control, and Dynamics*, Vol. 27, No.1, 2004, pp. 41-51.
17. Klyde, D. H., and C. Y. Liang, "Approach and Landing Flight Test Evaluation of Smart-Cue and Smart-Gain Concepts," AIAA 2008-6210, presented at *Atmospheric Flight Mechanics Conference*, Honolulu, Hawaii, 18-21 August, 2008.
18. Roscoe, A. H., and Ellis, G. A., *A Subjective Rating Scale for Assessing Workload in Flight. A Decade of Use*, Royal Aerospace Establishment Technical Report TR90019, Farnborough, U.K., 1990.