5-2014

A Quantitative Evaluation of Pilot-in-the-Loop Flying Tasks Using Power Frequency and NASA TLX Workload Assessment

Antonio Gemma Moré
University of Tennessee - Knoxville, amore@utsi.edu

To the Graduate Council:

I am submitting herewith a thesis written by Antonio Gemma Moré entitled "A Quantitative Evaluation of Pilot-in-the-Loop Flying Tasks Using Power Frequency and NASA TLX Workload Assessment." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Engineering Science.

Borja Martos, Major Professor

We have read this thesis and recommend its acceptance:

Peter Solies, Steve Brooks

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

A Quantitative Evaluation of Pilot-in-the-Loop Flying Tasks Using Power Frequency and NASA TLX Workload Assessment

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Antonio Gemma Moré

May 2014

DEDICATIONS

I dedicate the work that follows to my grandfather Anthony Enrico Gemma. Man walked on the Moon because of his accomplishments and I am honored to follow in his footsteps.

ACKNOWLEDGEMENTS

I would like to thank my family and friends for their support and encouragement throughout the pursuit of my Master of Science in Engineering Science at the University of Tennessee Space Institute. Accomplishments are never achieved without the help of others, and I owe a great debt to those around me who have guided my way. My mother, Kathleen Brenda Gemma, has always supported me and for that I am eternally grateful. For the last 23 years she has given me the strength to persevere and her dedication to our family is an example I hope to one day follow. No one could ask for a better brother than my brother, Marcos Gemma Moré, I am proud to see him graduate from Tennessee Technological University and begin a promising career in Mechanical Engineering. I want to also acknowledge my father Marcos Ortiz Moré, and grandparents Marcos Antonio Moré, Daisy Ortiz Moré, Anthony Enrico Gemma, Carmella Galinelli Gemma, and my uncle Anthony Thomas Gemma.

Additionally I would like to thank all my colleagues and mentors that have helped me along the way at UTSI. Special thanks go to Dr. Borja Martos, Dr. Peter Solies, and Dr. Steve Brooks for guiding my graduate education and aiding me in the completion of this thesis. I am humbled to have had the opportunity to learn from a group of such intelligent and intellectually honest men and I hope to represent them well in the future. Thanks must also be extended to Mr. Rich Ranaudo, Mr. Devon Simmons, Mr. Toby Sorensen, Mr. Jonathan Kolwyck, Mr. Greg Heatherly, and Mr. Jacob Bowman for their support and friendship throughout this process. Before I came to UTSI I only knew vague details about the sort of work going on “at the south end of the airport”, but after spending two years here I now fully recognize what a unique and fundamentally important team the University of Tennessee has assembled in Tullahoma.

Finally, I owe a great deal of gratitude to my wonderful girlfriend, Megan Carter, for her encouragement, patience, and thoughtfulness during this process. Thank you for being there for me.

ABSTRACT

While all manner of both qualitative and quantitative assessment tools exist to measure pilot performance during aircraft flight test, the argument to mathematically correlate two such diametrically different metrics is strong. By definitively connecting a pilot's written handling qualities or task loading feedback with measured performance data, researchers can more accurately examine any of a whole host of flight research topics.

Building upon past research which shows a positive correlation between Cooper-Harper Handling Qualities Ratings and calculated values for power frequency using a group of experienced test pilots, it is valuable to examine whether power frequency correlates with other metrics such as the NASA Task Loading Index (TLX). TLX provides a measure of a pilot's self-assessed workload and is routinely used in modern flight test experimentation to measure perceived pilot workload.

Using data from twenty-nine instructor pilots flying the NASA Ice Contamination Effects Flight Training Device (ICEFTD), the data set examined showed little connection between power frequency values and the TLX scores assigned by the pilots to each approach. Among the group of pilots flying the ICEFTD, self-assessed workload was a poor indicator of measured work load – such a trend indicates that non-test pilot self-measurement in workload assessment may not be as valuable as trained test pilot measurements. A number of influential causal factors were evident in the use of this recycled data set, and an ideal retest scenario is discussed at length.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

AGL	Above ground level
AIAA	American Institute of Aeronautics and Astronautics
AOA	Angle of attack
ATC	Air traffic control
ATP	Airline transport pilot
BAC	British Aerospace
BAR	Bihrle Applied Research
DA	Decision altitude
ERAU	Embry-Riddle Aeronautical University
FAF	Final approach fix
FFT	Fast Fourier Transform
FREDA	Frequency domain analysis
FTD	Flight training device
HQR	Handling qualities rating
ICEFTD	Ice Contamination Effects Flight Training Device
ICEPro	Icing Contamination Envelope Protection System
IMC	Instrument meteorological conditions
IPS	Icing Protection System
MATLAB ®	Matrix Laboratory
MFD	Multi-function display
NASA	National Aeronautics and Space Administration
NOAA	National Oceanographic and Atmospheric Administration
OFAT	One-Factor-At-A-Time
PIO	Pilot induced oscillation
PSD	Power spectral density
RAF	Royal Air Force
RMS	Root mean squares
STI	Systems Technology Incorporated
TIP	Tailplane Icing Program
TLX	Task load index
UTSI	University of Tennessee Space Institute

1. INTRODUCTION

1.1 PROBLEM STATEMENT

The University of Tennessee Space Institute's (UTSI) Aviation Systems program is tasked to perform airborne science missions and flight test research duties for a variety of governmental and nongovernmental customers including NOAA, NASA, and the Department of Defense. In support of that mission, UTSI personnel worked closely with Bihrle Applied Research (BAR) and were instrumental in the fundamental design, testing, and checkout of the NASA Icing Contamination Envelope Protection System (ICEPro) software package which was integrated into the NASA Ice Contamination Effects Flight Training Device (ICEFTD) simulator.

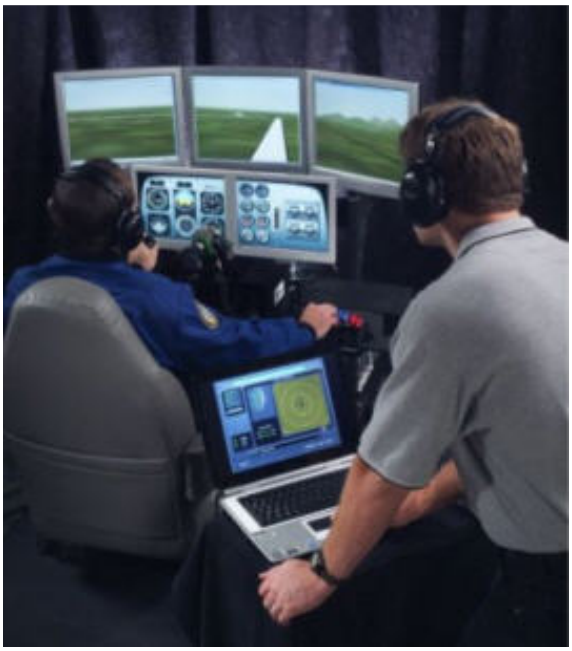

Figure 1: NASA Ice Contamination Effects Flight Training Device (ICEFTD) [1]

Figure 2: NASA DeHavilland "Twin Otter" Icing Research Aircraft [2]

The ICEFTD is a mobile simulator, shown in figure 1, which accurately models NASA's DeHavilland "Twin Otter" icing research aircraft, seen above in figure 2. Using the available icing database, the simulator can generate conditions ranging from a no-ice baseline configuration, to a failure of the icing protection system (IPS) following a 22.5 minute icing exposure, to a tailplane-only icing encounter [3]. Icing characteristics were collected through a combination of wind tunnel tests and in-flight data gathered using representative ice shapes which were fitted to the wing and elevator control surfaces and flown on the "Twin Otter" icing research aircraft.

The ICEPro software package uses algorithms to compare the baseline and iced performance models in order to provide the pilot safe-envelope airspeed, angle of attack, and flap extension envelope-limiting warning cues. Warnings were generated through real-time processing of the measured aircraft state, and the resulting stability and control derivatives were

then compared to a database in order to provide maximum and minimum airspeed and angle of attack cues. By operating the aircraft within the given envelope, ICEPro has been proven in simulation to provide effective real-time assessment cues which helped pilots avoid loss of control events.

In order to validate the ICEPro systems mission and to test its utility, 29 pilots were divided into a control group (using a baseline display) and an experimental group (using ICEPro) and flew identical precision approaches in simulated icing conditions [3]. In addition to the volume of aircraft performance data generated by the ICEFTD during each precision approach, pilots completed a NASA Task Load Index (TLX) questionnaire which rates perceived pilot workload using six broad subscales. A TLX score can range from 0 to 100, with 0 meaning a very low workload and 100 indicating a very high workload.

There are many tools at present to analyze aircraft handling qualities and pilot workload, chief of which are Cooper-Harper Handling Qualities Rating and the NASA TLX system, respectively. Good results have been found using experienced personnel flying very specific tasks with well defined performance parameters using Cooper-Harper and NASA TLX, but at best the feedback generated is qualitative in nature [4]. In order to quantify human performance in a simulator or aircraft it is highly desirable to analytically examine performance data instead of relying only upon pilot feedback.

In the past, crossover frequency has been used to analyze 'pilot in the loop' tasks with known forcing functions, but when pilot input itself is not known crossover frequency cannot be measured directly. In such a case cutoff frequency can be calculated and used, but it too has its flaws, chief of which is the inability to gage pilot intent. In other words, cutoff frequency is one-dimensional since it accounts for the magnitude of pilot activity ('what size amplitude

oscillations occurred' and 'how many oscillations occurred per unit of time') but it does not provide a measure of the level of pilot activity. Regions with low and high levels of pilot activity may record similar cutoff frequency scores, a problematic trend at the very least.

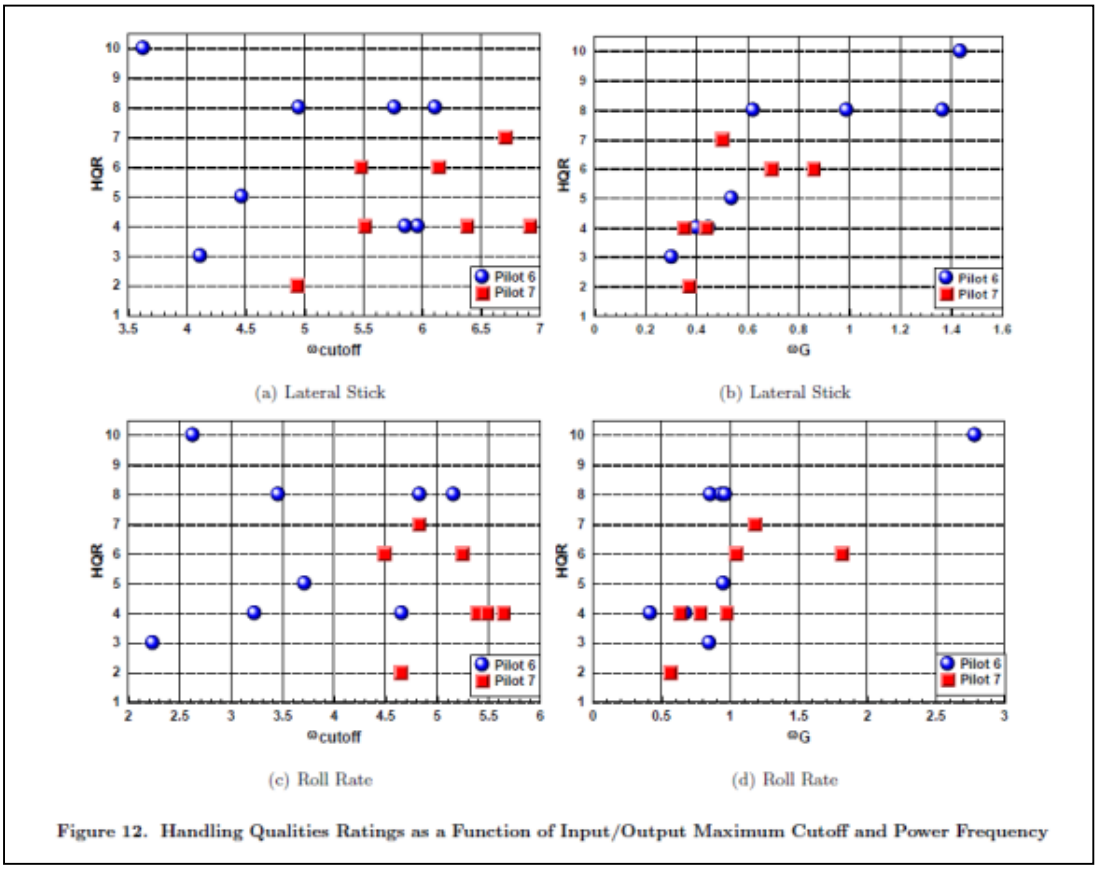

Figure 3: STI Learjet data - cutoff and power frequency vs. Cooper-Harper [4]

However, a new parameter known as power frequency has been shown to better tie the frequency of a pilot input with the corresponding intensity of that input. Researchers at Systems Technology Incorporated (STI) found a wide scatter of data using cutoff frequency versus Cooper-Harper HQR, visible above in plots a and c of figure 3. However, they also noted a

positive correlation between power frequency and Cooper-Harper HQR (plots b and d in figure 3) which was a first-of-its-kind connection at the time.

Using the aforementioned data collected in order to validate the ICEPro system aboard the ICEFTD, the following thesis seeks to compare power frequency to TLX scores in order to determine if a positive correlation exists. Since a calculation for power frequency can be performed for every single data point within a sample run, values for maximum and average power frequency were used and plotted against TLX scores in order to examine the question of causality. The method used is identical to the technique utilized by STI in their experimentation in 2009 [4].

2. FLIGHT TEST DATA

2.1 NASA 'ICEFTD' SIMULATOR DESCRIPTION

NASA has long been interested in gathering information on the issue of icing, and through the Tailplane Icing Program (TIP) a large quantity of data was collected [3]. As its name implies, TIP focused on the issue of ice buildup on an aircraft's horizontal tail, and through the use of artificial foam ice shapes, several tests were performed using the NASA DeHavilland "Twin Otter". A variety of maneuvers including wing flap transitions, airspeed sweeps, and engine power changes were performed and as a result NASA was able to note diminished longitudinal stability and elevator effectiveness for the 'iced' aircraft. Of note was a direct correlation between increasing flap angle and diminished longitudinal stability.

To better understand how pilots respond to icing and to develop a diagnostic teaching tool for future use, NASA applied the Ice Contamination Effects Flight Training Device, or "ICEFTD", to the problem. Built by Bihrle Applied Research, ICEFTD is a portable simulator system which models the cockpit of a Twin Otter airplane [1]. Using data gathered from the same wind tunnel tests and aircraft flight tests flown with ice shapes during the TIP program, the ICEFTD can provide an accurate representation of the Twin Otter aircraft in several configurations including a no-ice baseline and several different iced scenarios. The system can be programmed to simulate a gradual icing buildup, an "all on at once" configuration which models 22.5 minutes of icing exposure, or a tailplane-only icing setting.

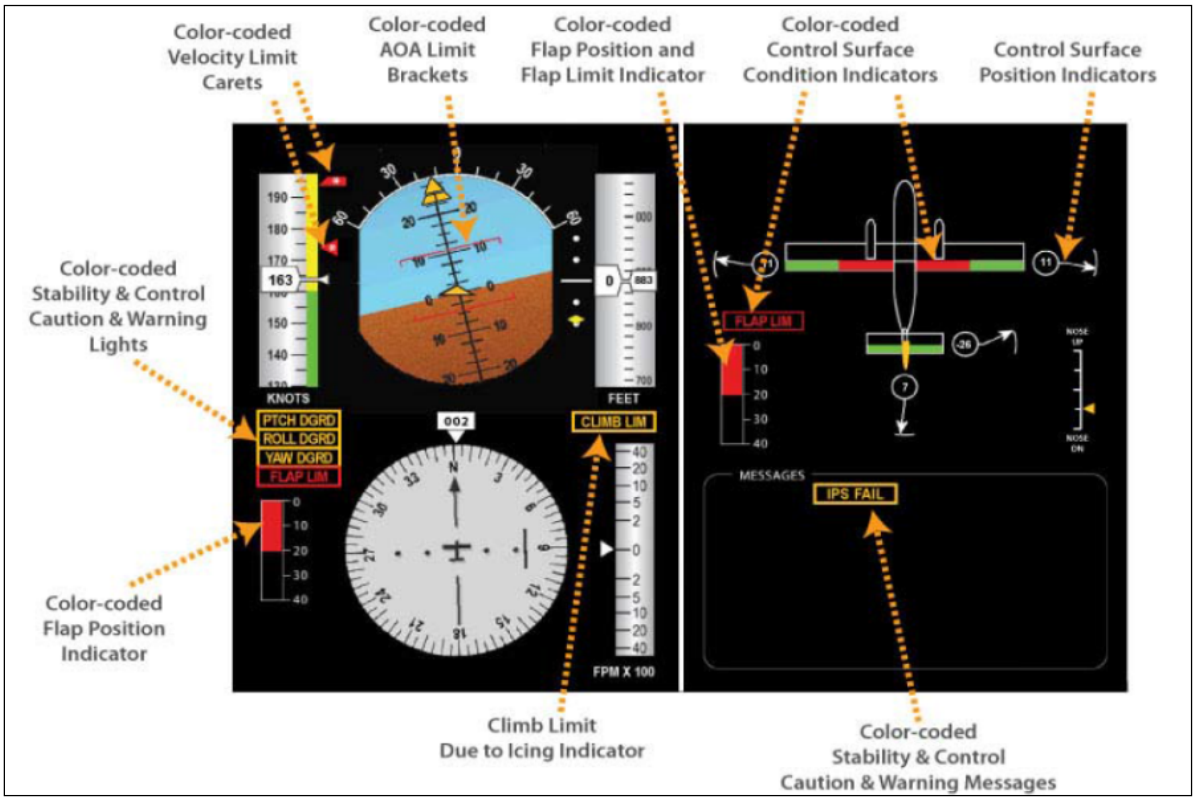

Figure 4: ICEPro flight display [3]

After the ICEFTD simulator and its associated icing database were certified, the ICEPro software package was integrated into and tested on the simulator. The ICEPro program replaced the simulators original cockpit display from a 'steam gauge' instrumentation panel to a modern multi-function display (MFD) style setup. Such a change aided in the ease of tying in ICEPro warning cues and was highly representative of both current and future aircraft cockpit display designs.

Shown above in figure 4, the basic ICEPro display highlights critical parameters as color-coded caution lights which only illuminate when the aircraft approaches the boundaries of an unsafe or unstable condition. Alerts for angle of attack, airspeed, climb performance, flap

position, and general performance degradation in all three axes are tied to their corresponding instruments.

For example, multicolored angle of attack (AOA) brackets illuminate on the pitch ladder to provide pitch limits for the pilot – the upper bracket indicates the predicted wing stall AOA, while the lower bracket is coupled to the minimum safe AOA to prevent a tail stall from occurring. The brackets are nominally white but will change to amber to indicate caution and red when an unsafe condition has been met. The intuitive approach is to maneuver the aircraft so as to remain within the AOA brackets, and by using the aforementioned cues which are tied to the aerodynamic model, a pilot is able to intuitively avoid regions of diminished performance due to icing contamination. Similarly actuated cues illuminate for airspeed and climb performance, while pitch/ roll/ yaw performance degradation alerts are instead accomplished via a series of text alerts on the right side of the MFD display. The flap position indicator simply turns red when an unsafe condition has occurred, prompting the pilot to retract wing flaps in order to return the aircraft to a stable state.

2.2 FLIGHT CONTROL SYSTEM DESCRIPTION

The ICEFTD was developed with the goal of providing a realistic training environment for pilot familiarization of typical effects of aircraft icing. The basic layout is made up of a metal framework underneath the pilot seat, a control yoke, rudder pedals, a throttle quadrant and flap control, and a series of computer monitors which display the ‘outside’ environment as well as the ‘internal’ cockpit instrumentation. Underneath the seat is a series of computers which control the simulation and displays and just behind the computer monitor array is a force feedback control loader.

Figure 5: NASA ICEFTD setup (without side curtains) [5]

Figure 5 above documents the ICEFTD setup which was used to gather data. Both the aileron and rudder flight controls are connected to simple springs which provide resistance and center the flight controls when no input is made. The only flight control which has force feedback is the elevator. The control loader connected to the elevator is capable of delivering over 150 lbs of force to pull or push the elevator away from the pilot [3]. The loader acted as a stick shaker and stick pusher when flight conditions dictated, simulating the feedback system present in commercial aircraft.

When in use, the ICEFTD was surrounded on all sides with a black curtain and lighting in the room was dimmed in order to minimize external distractions. In addition, the evaluation pilot and test conductor (located to the rear of the ICEFTD setup in figure 5) both wore aviation headsets and were connected together via an intercom system to add ATC-style communication

into the simulation. Finally the evaluation pilot was video and audio recorded using a digital video camera which looked over the pilot's right shoulder during each approach.

2.3 APPROACH AND LANDING TASK

In order to first build a baseline of relevant knowledge and build simulator proficiency, each pilot was given approximately 1.5 hours of training prior to data collection. The basic approach format was discussed in detail, and pilots flew in a no-ice baseline configuration in VMC and IMC conditions to conduct their practice approaches.

Since the original study being conducted was a validation of the ICEPro system architecture, pilots in the 'evaluation' group were given training using the ICEPro cockpit display and the 'baseline' group practiced with their steam-gage setup [3].

Each pilot performed three approaches during the evaluation phase of the test. Three personnel – the test subject, the test conductor, and the system operator – were involved with each test. The pilot sat enclosed inside the simulator cab which was surrounded with curtains to filter out distractions and extraneous information which might help or hurt the pilot's performance. No coaching or instruction was performed during each approach in order to preserve a sterile test environment.

Given minimal 'radar vector' style cues and standardized ATC-type commands by the test director, the pilot was directed to intercept the localizer/ glideslope in order to conduct a precision approach procedure. In accordance with the instructions, the pilot performed the approach, 'broke out' of IMC conditions approximately 400 feet above the ground (AGL), and continued their descent by transitioning to a visual approach. At 100' AGL the test director would order a missed approach procedure which entailed restoring full power and initiating a

climb. Once the pilot advanced both throttles the test conductor would then signal the system operator to fail an engine and the test would terminate when the test subject had turned the aircraft to the missed approach heading [3]. Both the test subject and test conductors were video and audio recorded, with the camera placement such that the test subject's face could not be seen by the camera.

Immediately after each evaluation run the test conductor would direct the test subject to fill out the electronic TLX form and then would provide a short debrief. The entire process of pre-test training, three runs for data, and accomplishing the necessary paperwork took approximately three hours per pilot, which was a limit imposed to prevent fatigue from affecting pilot performance. After the entire test was complete, the test subject filled out a post-test survey which sought to quantify the utility and overall assessment of the ICEPro system architecture – for obvious reasons those comments are not included here.

To simplify analysis of the general landing task, five individual portions of the approach were identified and analyzed. Using easily identifiable cues present in the data, all eighty approaches could be similarly deconstructed and examined in detail side by side.

While all five segments have importance and validity, two overarching themes were used for data analysis. First, an analysis of all five segments was performed. Since each individual segment of the approach may last for only a matter of 45 to 60 seconds (i.e. segment 1) there was less to be gained by fixating on short clips of the overall approach. In addition, most segments – with the notable exception of segment 5 – only involve a single maneuver or tracking task (segment 2 is simply a level 90º turn) which involve minimal control activity in multiple axes. Perhaps most importantly, the assigned TLX number reflects the pilots stated impression of the

entire approach and go-around sequence, so an examination of data for the full run best matches the pilots feedback.

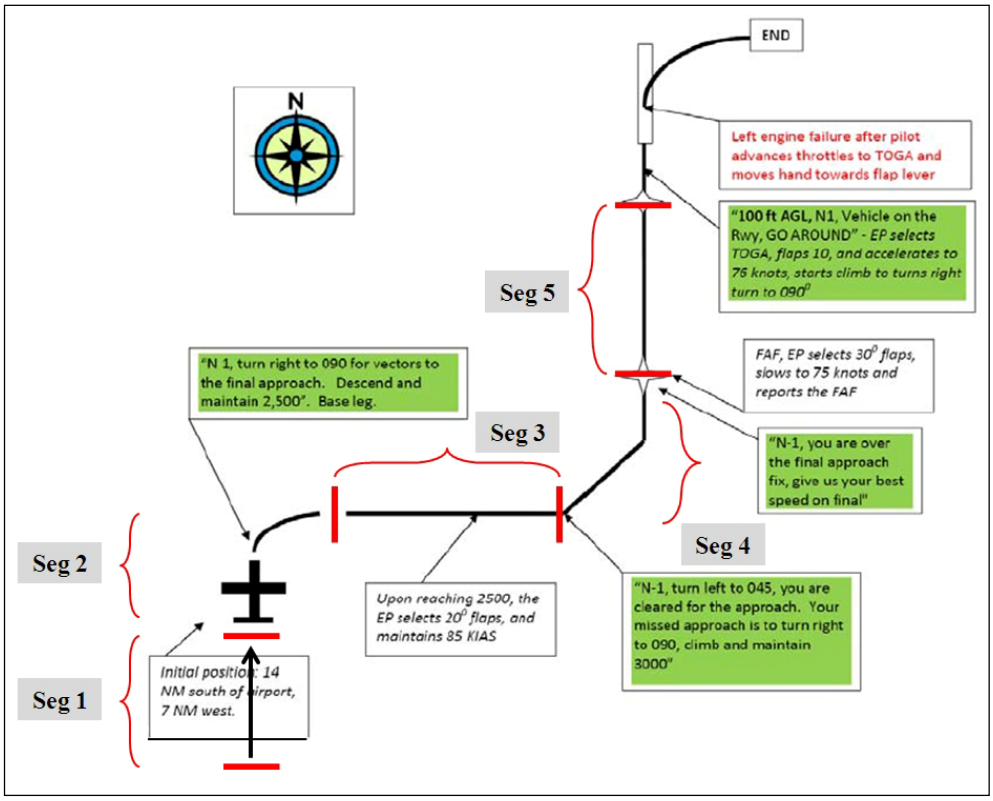

Figure 6: Approach and landing task outline, segments 1-5 [3]

The second approach used was to analyze Segment 5 data by itself. Since Segment 5 extends from the final approach fix (FAF) marker to the decision altitude (DA) marker, multiple control inputs occurred as the pilot strove to maintain ATP standards for airspeed, localizer, and glideslope while flying segment 5 of the approach. While segments 1 through 4 are generally considered low workload segments with few detailed performance parameters for a pilot to maintain at a time, segment 5 is the only section with detailed descent rate and course correction feedback. Most pilots took around four minutes to pass between the FAF to DA boundaries of

segment 5, which also served as the longest individual segment of the entire instrument approach task. Figure 6 above provides a detailed breakout of the five segments of the approach.

The twenty nine Embry-Riddle Aeronautical University (ERAU) volunteers were all instructor pilots who held FAA commercial pilots licenses with multi-engine and instrument ratings – none in the group had prior experience as test pilots or with the Cooper-Harper HQR scale [3]. The 'baseline' group of 14 was made up of 12 male pilots and 2 female pilots, while the 'evaluation' group of 15 had 12 males and 3 females. In terms of experience, the baseline group had a median of 1350 total hours and 122 multi-engine hours, and the evaluation group had a median of 1250 hours with a median of 100 multi-engine hours [3].

Three quarters of the pilots had no prior in-flight icing experience, while a full 89% did not feel that their "prior icing related knowledge/ experience would have prepared them for the test scenario". Over seventy percent felt that the NASA icing instructional video and online materials gave them a new appreciation and depth of understanding regarding aircraft icing.

3. WORKLOAD AND HANDLING QUALITIES ASSESSMENT TOOLS

3.1 NASA TASK LOADING INDEX

The NASA Task Load Index is a rating procedure which provides a workload score through the use of weighted averages [6]. A test subject provides feedback by noting their self-assessed performance using six major subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration.

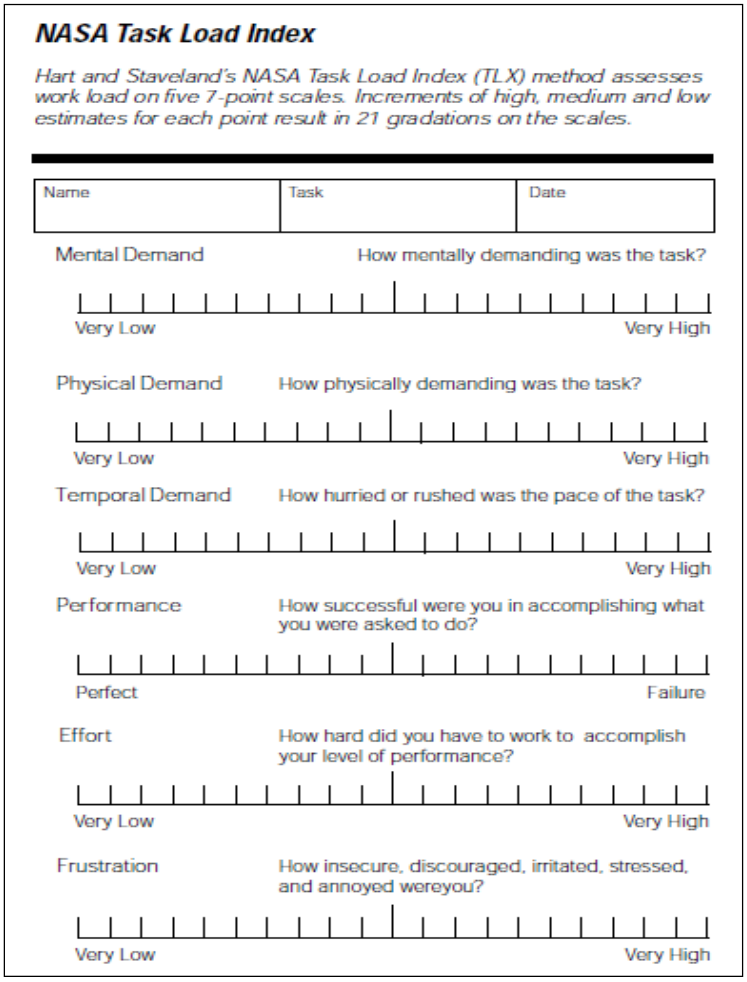

Figure 7: NASA Task Loading Index [7]

NASA TLX was generated after three years of work and over forty laboratory simulations and can either be performed as a "paper and pencil" test or digitally utilizing computerized calculations for speed. Since subjects can give feedback quickly it is possible to obtain a final rating rapidly in an operational setting. Another option for the researcher is to videotape the session and have the subject record their responses later, and through testing it was proven that "little information was lost when ratings were given retrospectively" [8].

Prior to conducting an evaluation, both the researcher and test subject are suggested to thoroughly familiarize themselves with the TLX instructions manual. The TLX evaluation itself is actually twofold: first, workload is divided into the six aforementioned subscales with 21 tick marks per scale ranging from 'very low' to 'very high'. The test subject then decides where their performance fell along the scale and marks it. Given this data, a rating is generated with a 0 to 100 range in 5 point increments. Used alone this value constitutes the "raw TLX" score, and the paper handout form used to calculate the raw score alone is seen in figure 7.

After completing the first part of the TLX procedure, the second portion of the TLX rating requires comparing the sources of workload. The six categories are compared to one another and the user selects which category was more relevant to workload. After summing up the number of times each particular subscale was chosen, a weighted score is found. Multiplying by the scale score for each dimension and then dividing by fifteen, a final workload score is created [8]. The final score can range from 0 to 100, with 0 meaning a very low workload and 100 indicating a very high workload. TLX has been used around the world to aid researchers in evaluating workload, and over 300 publications cite TLX scores as playing an integral part in their research [6].

3.2 COOPER-HARPER HANDLING QUALITIES RATING

The Cooper-Harper Handling Qualities Rating is perhaps the most ubiquitous measure of aircraft handling qualities and is still routinely used frequently by test pilots the world over. The current Cooper-Harper HQR was finalized in 1969 after several iterations and is a "broad strokes" decision-tree with a 1 to 10 scale where 1 is 'excellent' and 10 is 'major deficiencies'. By following the decision tree and arriving at a value, a pilot gives feedback on both performance – the precision of aircraft control attained by the pilot – as well as workload – the amount of effort and attention, both physical and mental, that the pilot must provide to attain a given level of performance [9].

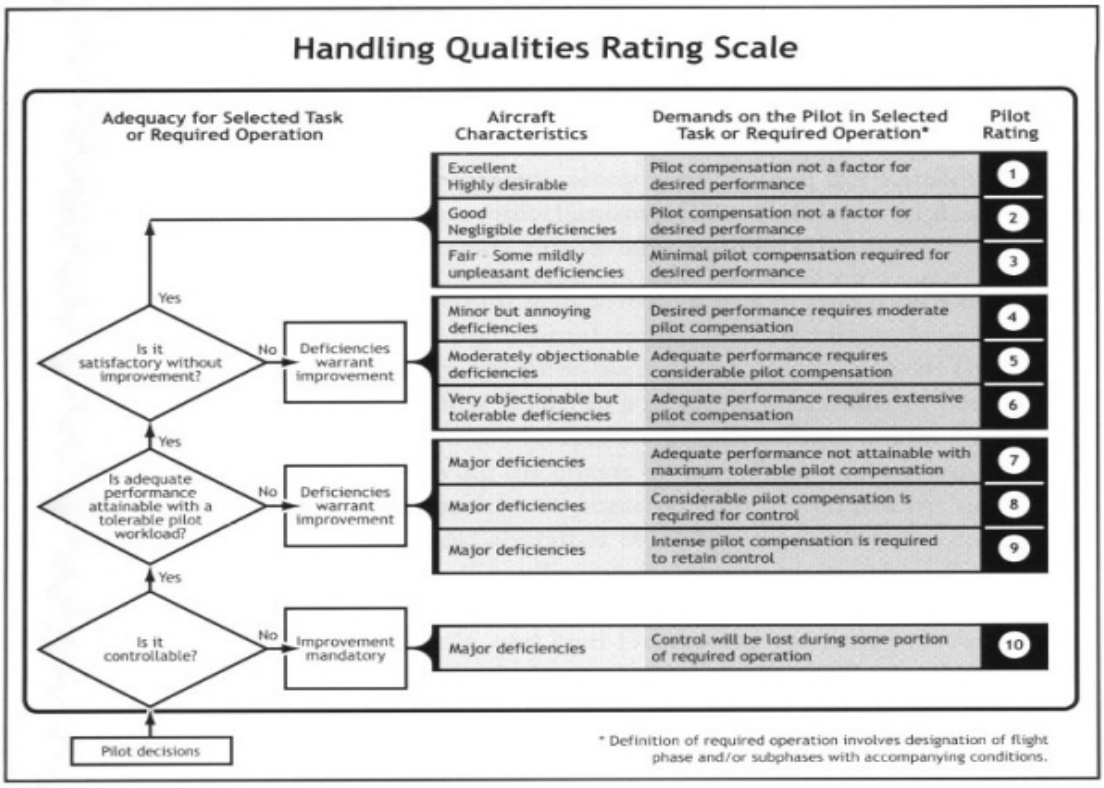

Figure 8: Cooper-Harper Handling Qualities decision tree [10]

While Cooper-Harper is still treated by many as the 'gold standard' for handling qualities evaluation, it still has its flaws. For one, the uni-dimensional format used in the Cooper-Harper method lacks diagnostic power and had been criticized for less than ideal reliability [11]. In addition, the new revised scale gives no provisions for failure considerations or emergency operations. Several tweaks have been made to the 1969 Cooper-Harper HQR, most notably a recent variant for assessing unmanned aerial vehicles [12].

In general, a low Cooper-Harper rating of 1 means handling qualities were described as "excellent/ highly desirable" where "pilot compensation not a factor for desired performance", as described in figure 8. In contrast, the highest Cooper-Harper rating of 10 indicates that "major deficiencies" exist where "control will be lost during some portion of required operations". As such, it is reasonable to expect that low Cooper-Harper values should correspond to low cutoff/ power frequencies while high Cooper-Harper values should correspond to high cutoff/ power frequencies. In other words, a pilot exerting low levels of feedback in order to attain the desired tolerances on the approach would likely rate the aircrafts handling qualities as close to ideal (i.e. 1) and vice versa.

Similar to a TLX workload score, a HQR cannot be broadly assigned to an aircraft or individual pilot. A HQR is a measure of an individual pilot's performance flying a well defined and repeatable task and can vary due to any number of conditions. Repeatability, consistency, and stability are the keys to meaningful analysis of a TLX score.

3.3 BEDFORD SCALE

Another decision tree workload assessment scale is the Bedford Scale. Due to a growing interest in accurately delineating pilot workload, the Royal Aircraft Establishment of Bedford

England developed the scale in the late 1960s to move from gathering pilot feedback to a more quantitative approach. After half a decade of work correlating pilot heart rate with pilot opinion of workload, researchers instead choose to attempt to define what pilot workload means and develop an improved means to measure it.

A.H. Roscoe and G.A. Ellis began their work by developing a questionnaire which was eventually distributed to over 350 military and civilian airline pilots [13]. After noting a great discrepancy in pilots' general understanding of workload – over 80% of the pilot's surveyed correlated workload with effort – Roscoe and Ellis tweaked the Cooper-Harper HQR definition of workload as follows: "pilot workload is the integrated mental and physical effort required to satisfy the perceived demands of a specified flight task" [13].

When tweaking the Bedford Scale, Roscoe and Ellis made great strides to ensure that each selection was adequately defined and highly specific. At first additional factors such as pilot fatigue were considered for inclusion, but after consulting with pilots it became apparent that additional factors overly complicated the process. In addition, since pilots generally prefer to compare their workload to some sort of 'baseline' (usually a previous experience), such a tendency complicated the task.

Decision Tree		Workload Description	Rating
Was workload satisfactory without reduction?	Yes	Workload insignificant.	1
		Workload low.	2
		Enough spare capacity for all desirable additional tasks.	3
	No	Insufficient spare capacity for easy attention to additional tasks.	4
		Reduced spare capacity. Additional tasks cannot be given the desired amount of attention.	5
		Little spare capacity. Level of effort allows little attention to additional tasks.	6
Yes			
Was workload tolerable for the task?	No	Very little spare capacity, but maintenance of effort in the primary task not in question.	7
		Very high workload with almost no spare capacity. Difficulty in maintaining level of effort.	8
		Extremely high workload. No spare capacity. Serious doubts as to ability to maintain level of effort.	9
Yes			
Was it possible to complete the task?	No	Task abandoned. Pilot unable to apply sufficient effort.	10

Figure 9: Bedford Scale [14]

As seen above in figure 9, the Bedford Scale is a ten rating scale ranging from 1 as 'workload insignificant' to 10 as 'pilot unable to apply sufficient effort'. A pilot proceeds from bottom to top first answering broad-based 'yes or no' questions and then moves on to more specific and descriptive options in order to assign a final rating. Although not explicitly noted on the Bedford Scale shown above in figure 9, so called 'half ratings' are permitted and are particularly helpful in evaluating lower workload tasks. As is the case with any workload task (or handling qualities task for that matter), great care must be taken to ensure that desired performance and standards are well defined in order to elicit accurate feedback.

The finalized version of the Bedford Scale was first put to the test by Harrier jump-jet pilots whilst using the ski-jump takeoff method [15]. In order to reduce takeoff distance aboard ships, an inclined 6º to 15º ramp was first fitted at RAF Bedford for trials. Pilots were asked to estimate handling qualities with Cooper-Harper and workload levels with Bedford during the test which involved starting from a full stop, accelerating down a short run, rotating the Harriers

nozzles rearward as the aircraft first passed over the edge of the ramp, and then transitioning into normal forward flight after launching off the ramp. Pilot heartbeat was also recorded for the eleven pilots to compare alongside their Bedford Scale feedback.

Workload ratings and heart rate data showed a positive correlation throughout the Harrier ski-jump test, and similar results were recorded for BAC 1-11 category 3 instrument approach and landing trials as well as during crew certification of the BAe 146 [15]. Pilots repeatedly found the scale "easy to use" and the only minor disagreements between Bedford Scale and heartbeat values were clearly attributable to pilot failure to rate the full period of the task.

3.4 PIO SCALE

From the Wright Flyer to the Lockheed Martin F-22, a seemingly simplistic source of pilot frustration has always come from pilot-induced oscillations, or PIOs. PIOs are frequently described as a "sustained or uncontrollable oscillations resulting from efforts of the pilot to control the aircraft" and are a result of the coupling of the aircrafts frequency with the frequency of the pilot's inputs [16]. While PIOs are often tangentially associated with a novice pilot applying excessive control inputs, PIOs can affect all types of aircraft and even high time aviators. A PIO is generally more severe on "short coupled" aircraft where the wing and tail surfaces are located closely together [16]. In order to effectively gage a PIOs severity and describe its tendencies, the PIO rating scale and their associated descriptions were developed.

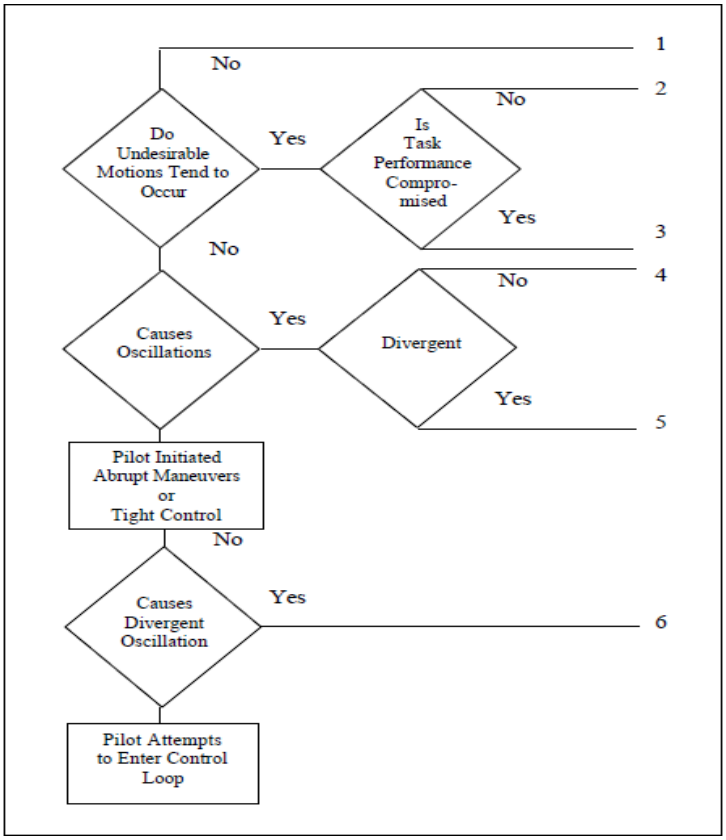

Figure 10: PIO Decision tree [17]

DESCRIPTION	NUMERICAL RATING
No tendency for pilot to induce undesirable motions.	1
Undesirable motions end to occur when pilot initiates abrupt maneuvers or attempts tight control. These motions can be prevented or eliminated by pilot technique.	2
Undesirable motions easily induced when pilot initiates abrupt maneuvers or attempts tight control. These motions can be prevented or eliminated but only at sacrifice to task performance or through considerable pilot attention and effort.	3
Oscillations tend to develop when pilot initiates abrupt maneuvers or attempts tight con-trol. Pilot must reduce gain or abandon task to recover.	4
Divergent oscillations tend to develop when pilot initiates abrupt maneuvers or attempts tight control. Pilot must open loop by relasing or freezing the stick.	5
Disturbance or normal pilot control may cause divergent oscillation. Pilot must open control loop by releasing or freezing the stick.	6

Figure 11: PIO rating scale descriptions [17]

Using the basic decision tree format (figure 10) a pilot starts at the bottom and answers basic yes or no questions. Corresponding PIO rating descriptions are outlined in figure 11 which allows a pilot to easier match the specific tendencies experienced to the appropriate rating. In contrast to the Cooper-Harper or Bedford Scale for handling qualities ratings, PIO ratings are based only on those broadly-based distinctions rather than tightly-worded descriptions.

4. CUTOFF FREQUENCY

4.1 CALCULATING CUTOFF FREQUENCY

In order to quantitatively assess a pilot's performance while flying a defined task, it is first necessary to define a method with which such an analysis can occur. At the lowest level, a time history of flight control and or surface deflections for aileron, elevator, rudder, flap, engine power, and any number of other flight controls provides a basic insight into the pilots control inputs.

A useful tool for analyzing pilot activity is cutoff frequency [4]. Cutoff frequency is a quantitative measure of pilot activity bandwidth in any control axis, and it is obtained by comparing the flight control input power to its frequency. Cutoff frequency provides a basic measure of the frequency of pilot activity (but not the level of that activity) in the form of a number. While imperfect, cutoff frequency does serve a purpose, and values for maximum cutoff frequency and average cutoff frequency were calculated for comparison purposes.

For example, low numbers for both maximum and average power and cutoff frequency means that the pilot is providing very small magnitude (input size) and small frequency (number of inputs per unit of time) inputs. In contrast, high numbers for both maximum and average cutoff and power frequency means that the pilot is providing very large magnitude (input size) and large frequency (number of inputs per unit of time) inputs.

Specifically, cutoff frequency is calculated by determining the frequency at which the integral of the power spectral density (PSD) – ranging from $\omega=0$ to $\omega=\infty$ – is half its total value. By calculating the ratio of root mean square (RMS) values and expressing them as ψ_1 / ψ_{total} a value for cutoff frequency is found. ψ_{total} can be calculated using equation 1.

$$\psi^2_{total} = \frac{1}{2\pi}\int_0^{\infty} G_{\delta\delta} d\omega \quad (1)$$

$$\psi^2_{1} = \frac{1}{2\pi}\int_0^{\omega_1} G_{\delta\delta} d\omega \quad (2)$$

In equation 1, the mean square value is essentially the area underneath the power spectral density (PSD) curve. Additionally, the PSD function for the controller is $G_{\delta\delta}$. Similarly to equation 1, ψ_1 is calculated in equation 2.

In equation 2 Ψ_1 is the root mean square value of the stick input signal over the frequency range $\omega=0$ to $\omega=\omega_1$. Since the critical value – cutoff frequency, or ω_{cutoff} – is at the half power point, the frequency where $\omega^2_{cutoff}/\psi^2_{total}=0.5$. This is also where $\psi_{cutoff}/\psi_{total}=0.707$ and where $\omega_1=\omega_{cutoff}$.

In order to rapidly calculate cutoff frequency Systems Technology Incorporated (STI) developed the FREquency Domain Analysis (FREDA) software [4]. FREDA is a Fast Fourier Transform (FFT) routine which identifies dynamic systems from flight test and simulation data. FREDA calculates the input PSD, output PSD, and remnant versus frequency in the process of identifying the system. Then, in order to calculate cutoff frequency a known pilot control inceptor input PSD is used although similar results may be obtained using the aircraft output PSD. The PSD is then numerically integrated and the frequency associated with the aforementioned half power is determined.

4.2 CALCULATING CUTOFF FREQUENCY AS A FUNCTION OF TIME

Since a relatively simple two-dimensional comparison of power versus time or frequency versus time can only yield so much information, a new way to illustrate the data was needed. Wavelet transforms plot power versus both time and frequency, and a specific type of wavelet

called scalograms are of interest and are frequently studied by STI. Since the wavelet transform time window decreases as frequency increases, a scalogram captures more details at higher frequencies than a corresponding Fourier transform.

$$\frac{\psi_1^2(t)}{\psi_{total}^2(t)} = \frac{\frac{1}{2\pi}\int_0^{\omega_1(t)} G_{\delta\delta}(t)d\omega}{\frac{1}{2\pi}\int_0^{\infty} G_{\delta\delta}(t)d\omega} = 0.5 \quad (3)$$

Cutoff frequency is calculated similarly to the process described above, and since integration over the frequency range is required, the time-varying cutoff frequency, $\omega_{cutoff}(t)$ is found by integrating the power over the frequency range for each time interval of the scalogram. The final equation for cutoff frequency is shown above in equation 3.

5. POWER FREQUENCY

5.1 CALCULATING POWER FREQUENCY

The chief drawback when examining cutoff frequency is its (relative) lack of depth regarding the magnitude of power compared to the power at all other times. The resulting time-varying cutoff frequency values can exhibit a variety of odd trends, most notably either fairly consistent figures or dramatic spikes at particular intervals in time. Due to the basic nature of the cutoff frequency calculation it is possible that areas with marginal amounts of control activity and areas with heavy activity will have the same cutoff frequency [4].

$$\omega_G(t) = \frac{\omega_{cutoff}(t)\, max G_{\delta\delta}(t)}{1000} \quad (4)$$

In order to move past cutoff frequency's chief limitation, a new parameter known as power frequency $\omega_G(t)$, was developed. Power frequency is found by multiplying the cutoff frequency at a particular time (t) by the maximum of the power spectral density, max $G_{\delta\delta}(t)$ over the frequency range, ω at time t. After then dividing by 1000, the value is scaled as seen in equation 4. By multiplying cutoff frequency by the maximum of the power spectral density, the rough tendencies of cutoff frequency are effectively tempered. As a result power frequency more accurately reflects pilot input.

In their work, researchers Amanda Lampton and David Klyde showed a positive correlation between pilot input power frequency and Cooper-Harper handling qualities value as shown in figure 3 [4]. The next logical step is to compare power frequency to NASA TLX workload assessment figures and determine if there is also a correlation.

6. ANALYSIS

6.1 NASA TLX VS POWER FREQUENCY

Using the 2009 data set, a wealth of relevant plots may be assembled. In order to make sense of these representations a little background is first necessary. All in all, eighty different approaches were conducted and several hundred individual parameters were recorded for the duration of the entire approach at a data collection rate of 10 Hertz [3]. Three parameters are of chief importance: "LATSTK" which is aileron deflection, "LONSTK" which is elevator deflection, and "RUD" which is rudder pedal deflection. Of these three primary variables, LONSTK is particularly important since in the ICEFTD pitch axis is the lone control surface with representative force feedback. Simple springs are connected to the aileron and rudder flight controls which allow those axes to mimic their corresponding control surfaces, but not match their performance entirely.

The driving question behind this research paper is to determine whether either a positive, negative, or correlation exists between power frequency and TLX: as such, the plots of value were clearly power frequency versus TLX score. Plots of power frequency versus TLX were generated for all three control axes, including both maximum and average values of power frequency, and for the entire approach or just for segment 5. For comparison purposes an identical series of plots examining cutoff frequency in the place of power frequency were also prepared.

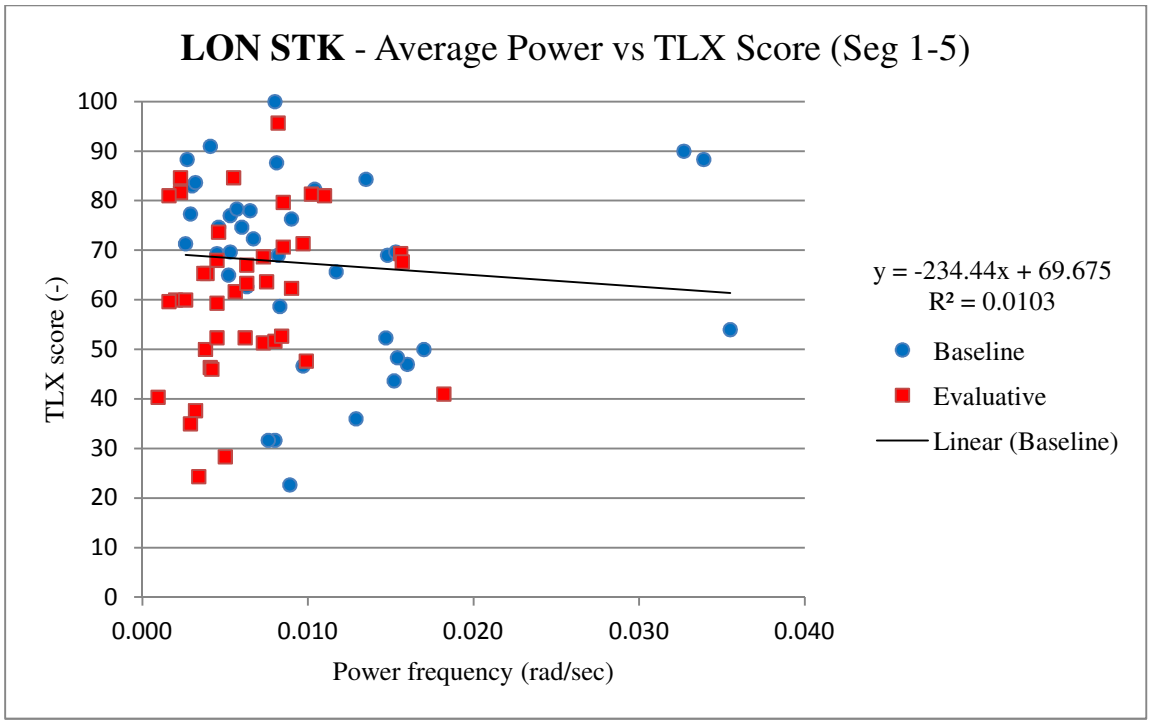

Figure 12: Whole run, average elevator power frequency vs. TLX

Were a distinct trend to be present between power frequency and TLX score it makes sense that it would first appear in figure 12 above. Examining the plotted values it is clear that a wide array of TLX scores – from 24.3 to 95.6 – gives a wide variety of perceived pilot workload to analyze. LONSTK average power values are fairly tightly clumped together between 0.002 and 0.014 radians per second with several outliers at around 0.035.

Disappointingly there is no easily identifiable linear (or nonlinear) trend between the average power values and TLX values, as all the data are grouped fairly close together. The average power frequency – a measure of rotational speed – appears to be clustered around 0.006 radians per second which is equivalent to 0.0009 Hz or roughly 0.34 degrees per second. While these values are small, compared to measured aileron and rudder activity they are relatively large, especially considering that control loading is only present in the LONSTK axis.

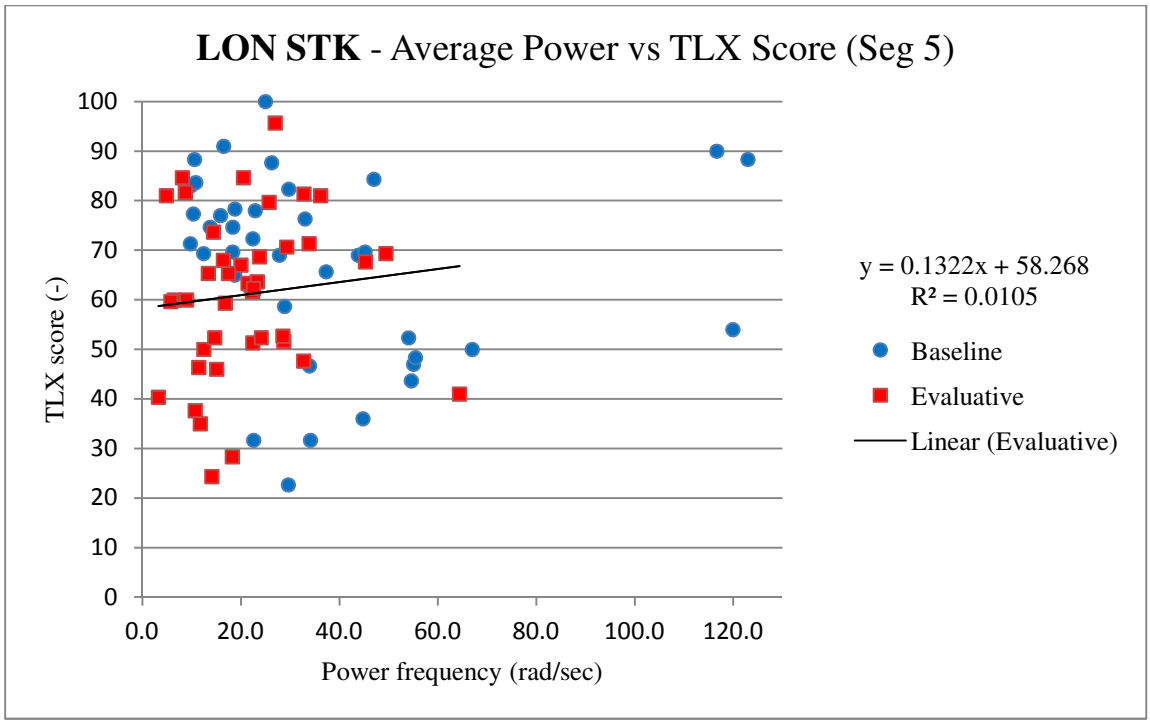

Figure 13: Segment 5, average elevator power frequency vs. TLX

An alternate strategy is to examine the total approach on a segment by segment basis. By switching from a macro to a micro scale it is possible to eliminate shorter portions of the approach with minimal control activity and instead focus on segments of greater interest. Since segment 5 is defined as the region between the final approach fix and decision altitude, it makes sense that pilots would be higher gain while they attempt to follow glideslope and localizer cues. Figure 13 above shows segment 5 values for average power frequency compared to their corresponding TLX score, and just like the prior plot a wide scatter is present with no definitive correlation visible.

Though the plots are not included here (see appendix 1 and 2), nearly identical results were found when both average and maximum power frequency values were plotted for aileron and rudder inputs. Since control loading was present in the LONSTK axis it makes sense that

any possible correlation should present itself there and possibly be reflected throughout the aileron and rudder data. At no point was a clear correlation present between average or maximum values of power frequency and the TLX values for aileron, elevator, or rudder performance.

6.2 SIX CRITICAL CASES

In order to provide a more detailed look at the general trends observed by plotting power frequency versus TLX, six critical cases were identified. As figure 14 below shows, the approaches with both the highest and lowest TLX values, number of tail stalls, and airspeed Theil inequality coefficients are all of interest. It is interesting to observe that five of the six critical cases are evaluation data runs, while only one was a baseline approach conducted without the aid of the ICEPro system display.

Point of interest:	Pilot:	Run type:	Run number:	Value of interest:
High TLX number	13	E	02	95.6667
Low TLX number	11	E	02	24.3333
High number of tail stalls	24	B	02	24
Low number of tail stalls	15	E	03	0
Highest airspeed Theil coefficient	13	E	01	0.52
Lowest airspeed Theil coefficient	20	E	01	0.17

Figure 14: Six critical cases

CASE 1: High TLX Score (13 E 01)

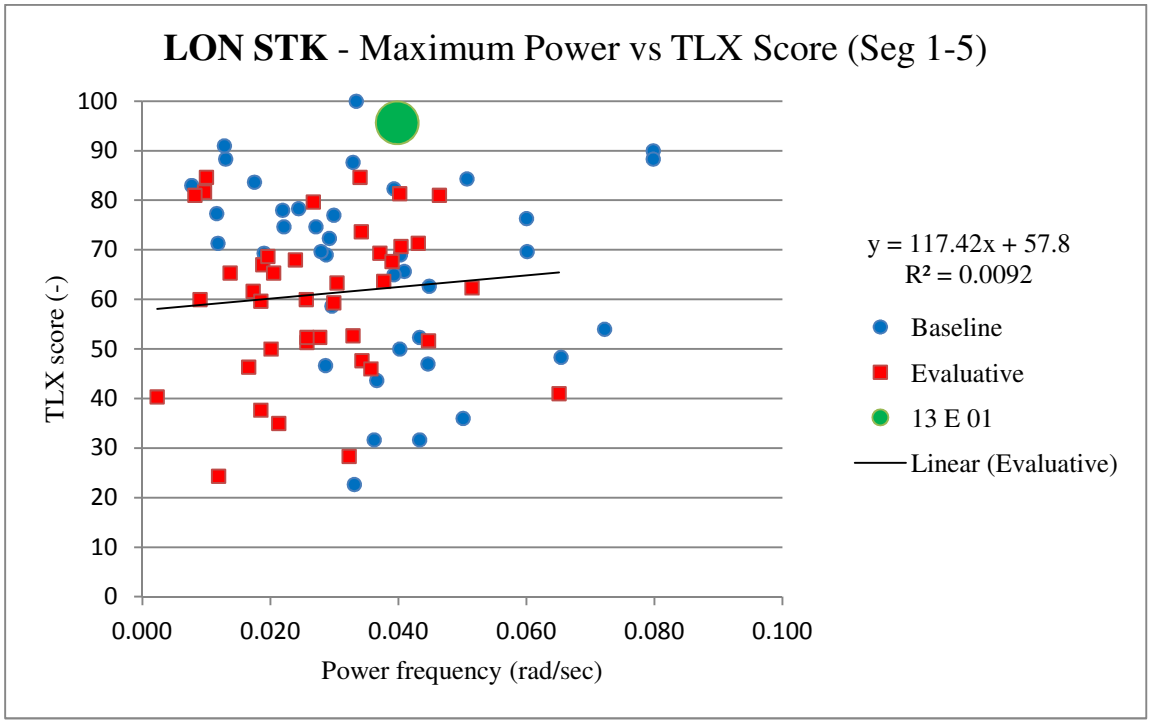

Figure 15: Highest TLX value; maximum elevator power frequency vs. TLX

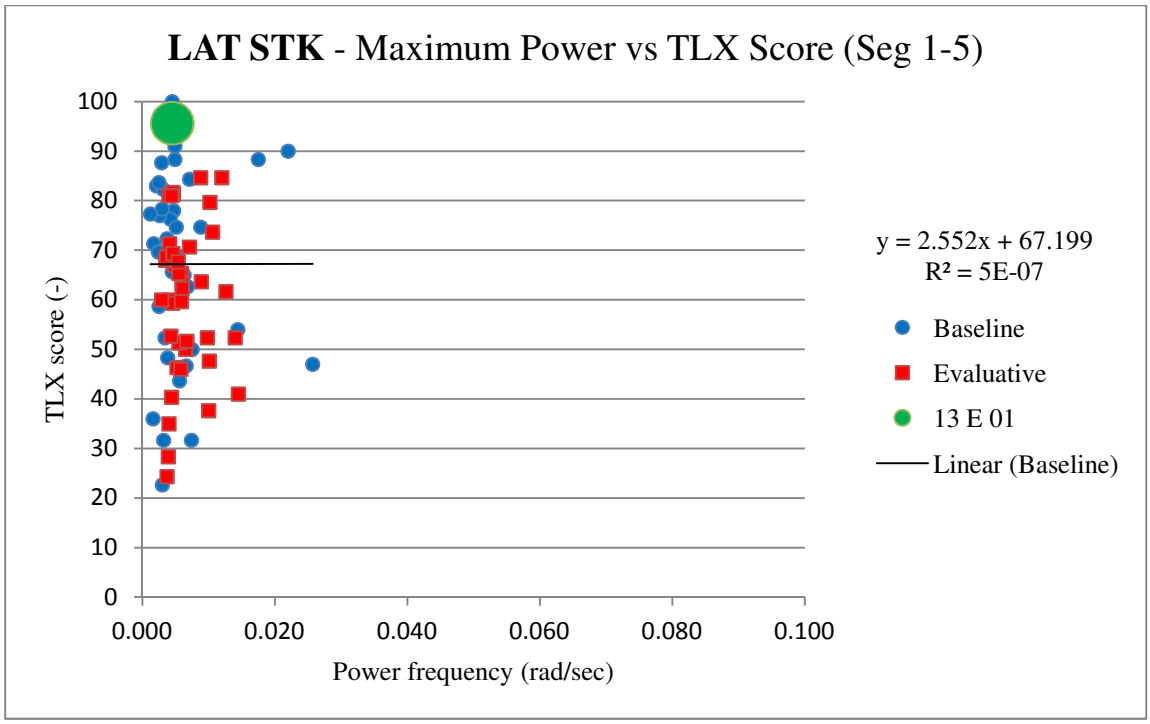

Figure 16: Highest TLX value; maximum aileron power frequency vs. TLX

As the title entails, pilot 13 flying his first of three approaches recorded the highest task loading index of any pilot flying any approach throughout the entire test. Figure 15 above illustrates that pilot 13's high TLX score was among the highest half of power frequency values recorded, but to draw a distinct correlation between high workload levels and high power frequency values for only one pilot is challenging.

While the elevator axis – outfitted with control loading – is the most likely region to expect a relationship to present itself, the highest TLX score approach does not identify any trends when examining aileron and rudder inputs either. Highlighted in detail in appendix 1 and above in part for the aileron axis in figure 16, aileron and rudder inputs, as a whole, occur at far lower frequencies that elevator inputs, but even so no distinctive pattern occurs. Again as was shown for elevator power frequency, the aileron power frequency is in the middle of the pack

with respect to the overall set of data. Of note is the observation that cutoff frequency values for aileron and rudder show higher values corresponding with the highest TLX score (appendix 1).

CASE 2: Low TLX Score (11 E 02)

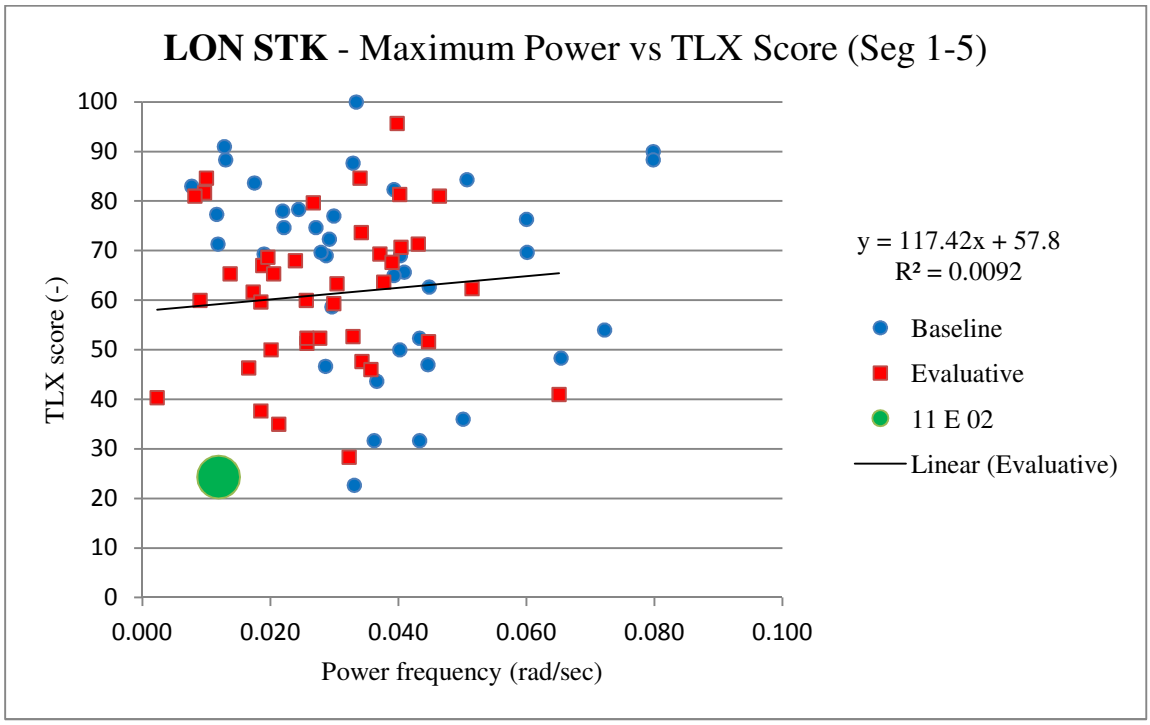

Figure 17: Lowest TLX value; maximum elevator power frequency vs. TLX

In direct contrast to case 1, Pilot '11E' recorded the lowest TLX score for the second approach of three, making this the run where the pilot felt the lowest level of workload for any run during the entire test. As such, with such a low TLX workload score it is reasonable to expect low values of power frequency as the pilot indicated through their feedback that their workload was small and they were able to maintain the desired tolerances with ease.

When looking at both plots for maximum (figure 17 above) and average power frequency versus TLX, the aforementioned expected trend does in fact bear itself out. Pilot 11 generated

some of the lowest power frequency values recorded throughout the entire study, and also some of the lowest TLX scores in general, and such results make logical sense. Run 11 E 02 confirms the projected hypothesis and connects to the observation made in case 1, but even the two cases taken together cannot fully substantiate a positive correlation for the study.

CASE 3: Highest number of tail stalls (24 B 02)

Cases 3 and 4 introduce a new metric for measuring pilot performance: the number of tail stalls which occurred per run. A particular characteristic inherent in the DeHavilland "Twin Otter" in icing conditions is a higher than normal proclivity to tail stalls as compared to other comparable aircraft. While an aerodynamic stall can occur on any aerodynamic surface – wing stalls are commonly associated with 'stalling' an aircraft, but tail stalls or even a rudder stalls can occur – tail stalls are a significant hazard for the Twin Otter.

As the original 2009 data was collected chiefly to determine the utility of the ICEPRO software package in aiding a pilot in icing conditions, pilots were specifically briefed on tail stalls and special care was taken to look for their occurrence in the data. While an aerodynamic stall occurs when the airfoil meets or exceeds is critical angle of attack, wing stalls are usually associated with a 'pitch up' in order to reach the critical AOA. In contrast, an icing-induced tail stall occurs at a critical airspeed and often occurs in the nose-down phase of flight. In broad terms, a pilot who experienced a number of tail stalls likely flew his or her approach at the upper-end of the prescribed airspeed range and thus encountered tail stalls frequently along the approach. A number of tail stalls is synonymous with a poorly flown approach, while no tail stalls indicates a well flown approach.

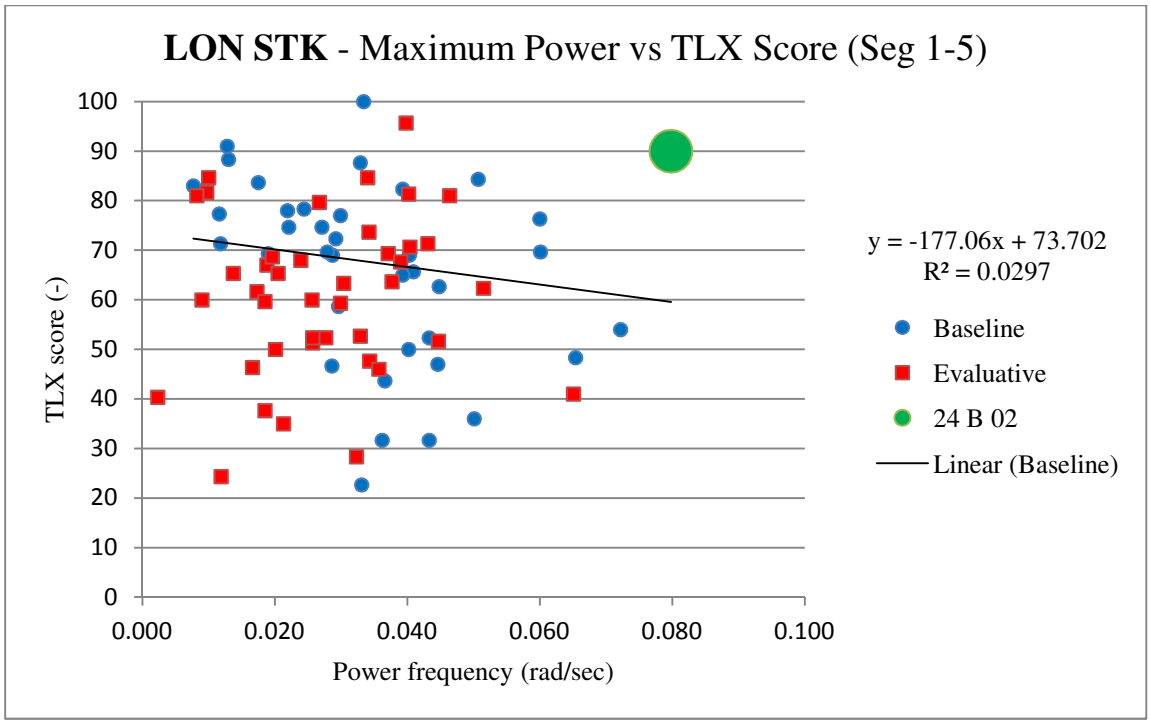

Figure 18: Highest number of tail stalls; maximum elevator power frequency vs. TLX

Case 3 examines the second baseline run of Pilot 24 where 24 tail stalls occurred. Examining a time history of the flight, tail stalls occurred throughout the entire approach but mostly within the last 400 seconds of flight which roughly corresponds to segment 5. Coupled with the high number of tail stalls are a great number of stick shaker warnings which were programmed into the simulator environment to provide an additional warning as the aircraft approaches a stall. Figure 18 above also confirms that run 24 B 02 was also the approach with the highest power frequency value of all eighty approaches.

Pilot 24 is the lone 'baseline' pilot of the six critical cases who exhibited outlier tendencies – while such a characteristic is not necessarily indicative of a larger tendency, it is at a minimum noteworthy. Pilot 24 also recorded three very high TLX scores - 84, 90, and 88 - which indicate a high perceived pilot workload.

CASE 4: Lowest number of tail stalls (15 E 03)

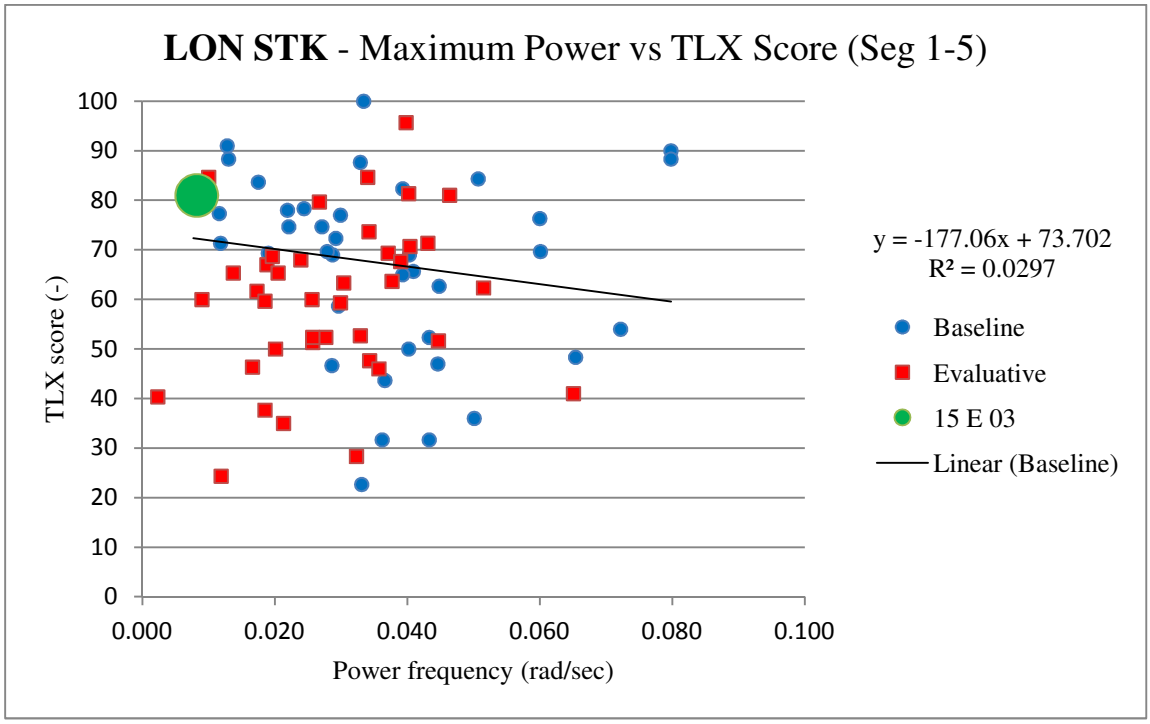

Figure 19: Lowest number of tail stalls; maximum elevator power frequency vs. TLX

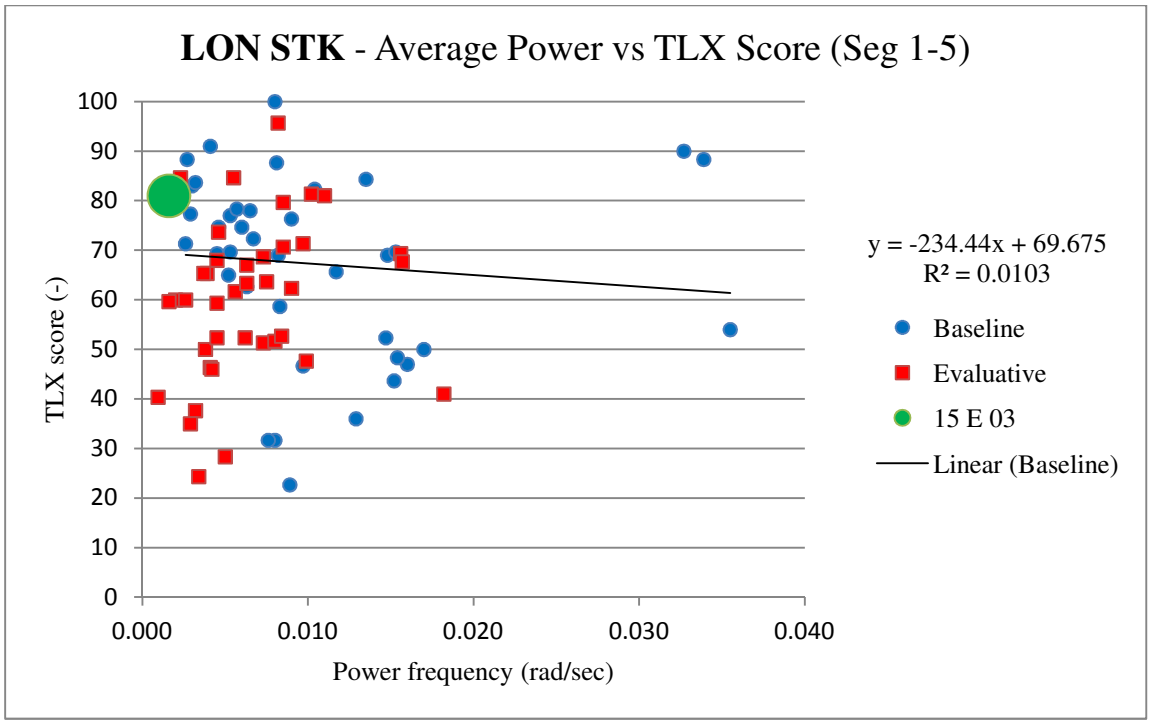

Figure 20: Lowest number of tail stalls; average elevator power frequency vs. TLX

Pilot 15s third run recorded the lowest number of tail stall of any approach. Coincidently, run 15 E 03 were also one of the runs with the lowest average and maximum power frequency values of all the approaches (see figures 19 and 20). Interestingly, pilot 15 recorded a TLX score of 81 for run #3, indicating the pilot felt a high level of workload was necessary in order to fly a tightly coupled approach within tolerances, all while avoiding tail stalls. While pilot 24 (case 3) understandably also had a high TLX score, such a trend here is surprising since it is natural to expect a pilot who had such low average power frequency to also have a low workload score.

CASE 5: Highest airspeed Theil inequality coefficient (13 E 01)

While there are a number of ways to quantify the accuracy with which an instrument approach was flown, measuring the ability to remain on glideslope, on localizer, and on airspeed are very important. When observing the cockpit instrumentation from the pilots perspective it is simplest to note 'on airspeed' (or operating within a desired +/- range from that airspeed) or gage localizer and glideslope performance by stating how many dots (a calibrated measurement on the instrument's face) far away from 'perfect' the aircraft is. Clearly it is valuable to condense such diverse parameters down to an easily definable single value for analysis, and a calculated Theil inequality coefficient does just that.

Using the ICEFTD simulator environment it is possible to measure the simulated Twin Otter's deviation from the desired flight path, and airspeed measurement is a simple comparison between the desired and measured airspeed numbers. By performing a fairly simple series of slope calculations, a Theil value provides a measurement of how far from the desired value a measured performance value is. A Theil of 0 is a perfect match with the desired tolerance while a Theil of 1is no match at all.

Case 5 is the highest airspeed Theil of all the runs, a 0.52. During the run pilot 13 had a high TLX score of 95.6 – note that the run identified as the highest airspeed Theil value (case 5) is also examined in case 1 as the highest TLX workload assessment number. Run 13 E 01 represented the highest workload and the worst airspeed performance of the entire test.

CASE 6: Lowest airspeed Theil inequality coefficient (20 E 01)

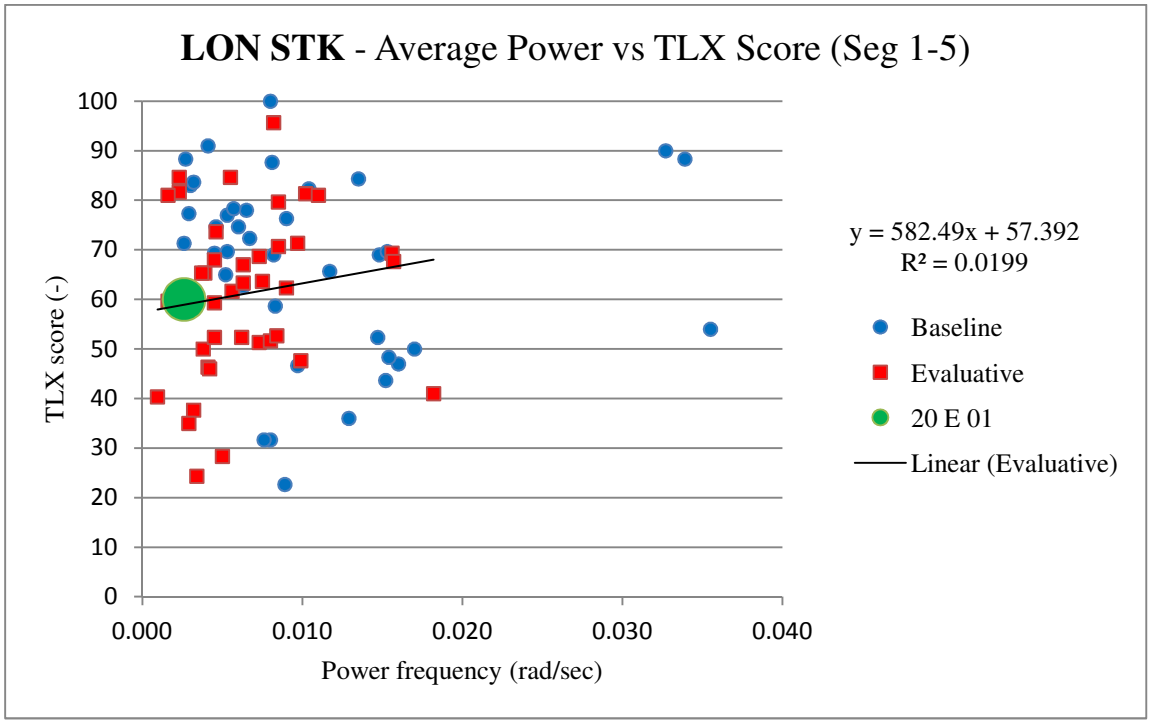

Figure 21: Lowest airspeed Theil coefficient; average elevator power frequency vs. TLX

The final run of interest is 20 E 01 where the pilot recorded an airspeed Theil inequality coefficient value of just 0.17 – pilot 20 did the best job of all twenty nine aviators evaluate during eighty runs in maintaining the desired airspeeds during all phases of the approach. Tied to this exemplary airspeed management were a low average power frequency of just 0.0026 rad/sec (shown above in figure 21) and a maximum power frequency of 0.025 rad/sec.

Connected with such low power frequency figures, pilot 20 also recorded a TLX workload score of 60 which is high given the ideal airspeed trend. Pilot 20's TLX scores declined slightly throughout his three approaches (60, 59.6, and 40.3) while both average and maximum power frequency values declined as well. Run 20 E 01 aileron and rudder power

frequency numbers were also very low, but there were seven stick shaker events in the go-around region of the approach along with a tail stall event.

7. IDEAL RETEST

While STI research comparing power frequency values to handling qualities ratings yielded a positive correlation, no such easily identifiable trend exists for the eighty approaches examined here when comparing power frequency with workload assessment. Even though the current data set is certainly valid, several key points make it less than ideal for examining the power frequency/ TLX relationship question. Moving forward an 'ideal retest' must be performed in order to better answer the question asked here; section 7 addresses that retest scenario.

7.1 PRIOR EXPERIMENTATION SETUP

In order to better understand the test environment used in the original study, it is valuable to examine the experiment methodology used by STI to perform their 2011 power frequency research. Researchers Amanda Lampton and David Klyde sought to examine what relation, if any, existed and a direct correlation between Cooper-Harper handling quality values and power frequency emerged [4].

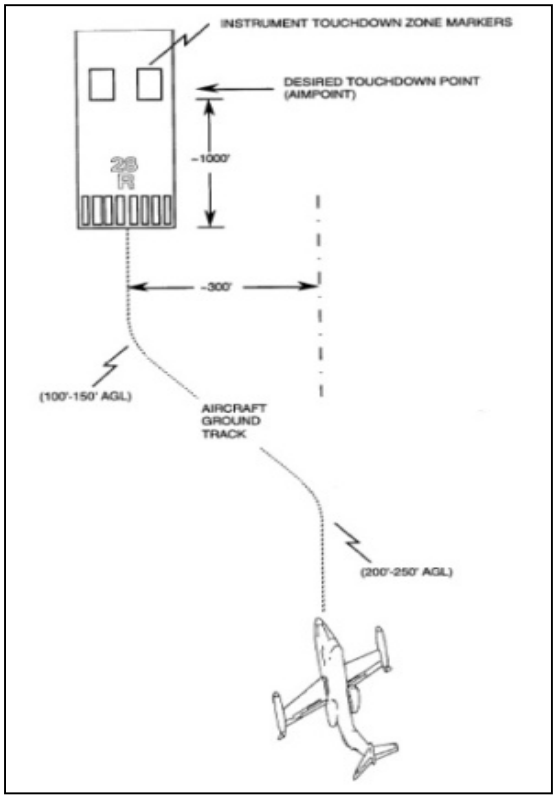

Figure 22: Precision offset landing task [4]

Lampton and Klyde performed a real world flight test aboard the Calspan Learjet 25 "in-flight simulator" to gather their data. Using two experienced test pilots each flying 11 and 8 evaluation runs, respectively, a precision offset landing task was performed. By displacing the pilot approximately 300 feet to the right of the extended centerline on final approach before tasking the pilot with aggressively returning to a proper heading and glide path for landing, the offset landing task is a standardized high-gain maneuver conducted at low altitude in the runway landing environment. Figure 22 above illustrates the basic maneuver in detail.

Since all nineteen STI tests were performed aboard a real aircraft in flight, the very real possibility of a varying atmosphere plays a factor in the test. "Similar weather conditions" existed during all test runs evaluated by the researchers, and various configurations were dialed

into the variable stability aircraft [4]. The test pilots evaluated each task immediately after completion using both a Cooper-Harper HQR sheet and the PIO scale. Given that Lampton and Klyde were able to successfully attribute power frequency calculations to Cooper-Harper HQR ratings, it would be foolish not to at least emulate their standard of testing, all the while expanding the scale of the experiment when and where it is appropriate.

7.2 DESIGN OF EXPERIMENTS

2009 ERAU Data

Pilot	Run	Type
1	1	Baseline
1	2	Baseline
1	3	Baseline
2	1	Evaluative
4	1	Evaluative
4	2	Evaluative
4	3	Evaluative
5	1	Evaluative
5	2	Evaluative
5	3	Evaluative
6	1	Evaluative
6	2	Evaluative
6	3	Evaluative
7	1	Baseline
7	2	Baseline
7	3	Baseline
8	1	Baseline
8	2	Baseline
8	3	Baseline
9	1	Baseline
9	2	Baseline
9	3	Baseline
10	2	Evaluative
10	3	Evaluative
11	1	Evaluative
11	2	Evaluative
11	3	Evaluative
12	1	Baseline
12	2	Baseline
12	3	Baseline
13	1	Evaluative
13	2	Evaluative
13	4	Evaluative
14	1	Baseline
14	2	Baseline
14	3	Baseline
15	1	Evaluative
15	2	Evaluative
15	3	Evaluative
16	1	Baseline
16	2	Baseline
16	3	Baseline
17	1	Evaluative
17	2	Evaluative
17	3	Evaluative
18	1	Baseline
18	2	Baseline
18	3	Baseline
19	1	Evaluative
19	2	Evaluative
19	3	Evaluative
20	1	Evaluative
20	2	Evaluative
20	3	Evaluative
21	1	Evaluative
21	2	Evaluative
21	3	Evaluative
22	1	Baseline
22	2	Baseline
22	3	Baseline
23	1	Evaluative
23	2	Evaluative
23	3	Evaluative
24	1	Baseline
24	2	Baseline
24	3	Baseline
25	2	Evaluative
25	3	Evaluative
26	1	Baseline
26	2	Baseline
26	3	Baseline
27	1	Evaluative
27	2	Evaluative
27	3	Evaluative
28	1	Baseline
28	2	Baseline
28	3	Baseline
29	1	Baseline
29	2	Baseline
29	3	Baseline

Figure 23: Summary of ICEPro validation runs

The data set from Embry-Riddle of 29 pilots was fairly evenly split, with 13 pilots flying the simulator as ‘baseline’ pilots without the ICEPro display and 15 flying as ‘evaluative’ pilots who flew with the ICEPro displays. While the initial design was for all pilots to conduct three

approaches each, due to several factors one pilot recorded only one run, two pilots flew two runs, and the rest flew all three as summarized in figure 23.

STI Learjet data

Pilot	Run	Cue?	Gain	Friction	Gradient
1	04	No			
1	05	No			
1	06	Yes	X		
1	07	Yes	X		
1	08	Yes	X		
1	09	Yes	X	X	
1	10	Yes	X	X	
1	11	Yes	X	X	X
1	12	Yes	X	X	X
1	13	Yes	X	X	
1	14	Yes	X	X	X
2	12	No			
2	13	No			
2	14	Yes	X		
2	15	Yes	X	X	
2	16	Yes	X	X	
2	17	Yes	X	X	X
2	18	Yes	X	X	X
2	19	No			

Figure 24: Summary of STI power frequency/ CH runs

The STI researchers selected data from two experienced test pilots but each pilot performed multiple runs while control gain configurations were altered – specific configuration changes were not disclosed [4]. By flying with only two test pilots, STI potentially limited the variance of data seen when using a large number of evaluative pilots. In contrast, they also ensured that both pilots provided informed Cooper-Harper HQR feedback due to their familiarity through repetition with the aircraft and the experiment setup. Between both pilots nineteen approach tasks were conducted, seen above in figure 24.

Since an ideal retest comparing power frequency to TLX does not strive to validate the ICEPro system like in the data set examined here, a different number of approaches and evaluative pilots must be considered. For reasons explained in section 7.5 it is valuable to split

the group into two equally sized elements, so proceeding forward with that assumption an analysis can be performed.

While a large volume of work has been performed in the field of design of experiments, definitively determining sample size and defining an appropriate confidence level remains a continuing challenge. However, work performed by Jakob Nielsen of the Nielsen-Norman Group provides a basic reference which is useful here [18]. Mr. Nielsen suggests a sample size of 20 test subjects for a quantitative study for several reasons, chief of which is an acceptable confidence interval. To begin, around 6% of data in any such test is likely an outlier (a figure calculated through and reinforced by experimental testing), so by removing the data from one of the 20 pilots a 'true' sample size of 19 is found – this leaves a +/-19% margin of error for a group of 19 users [18].

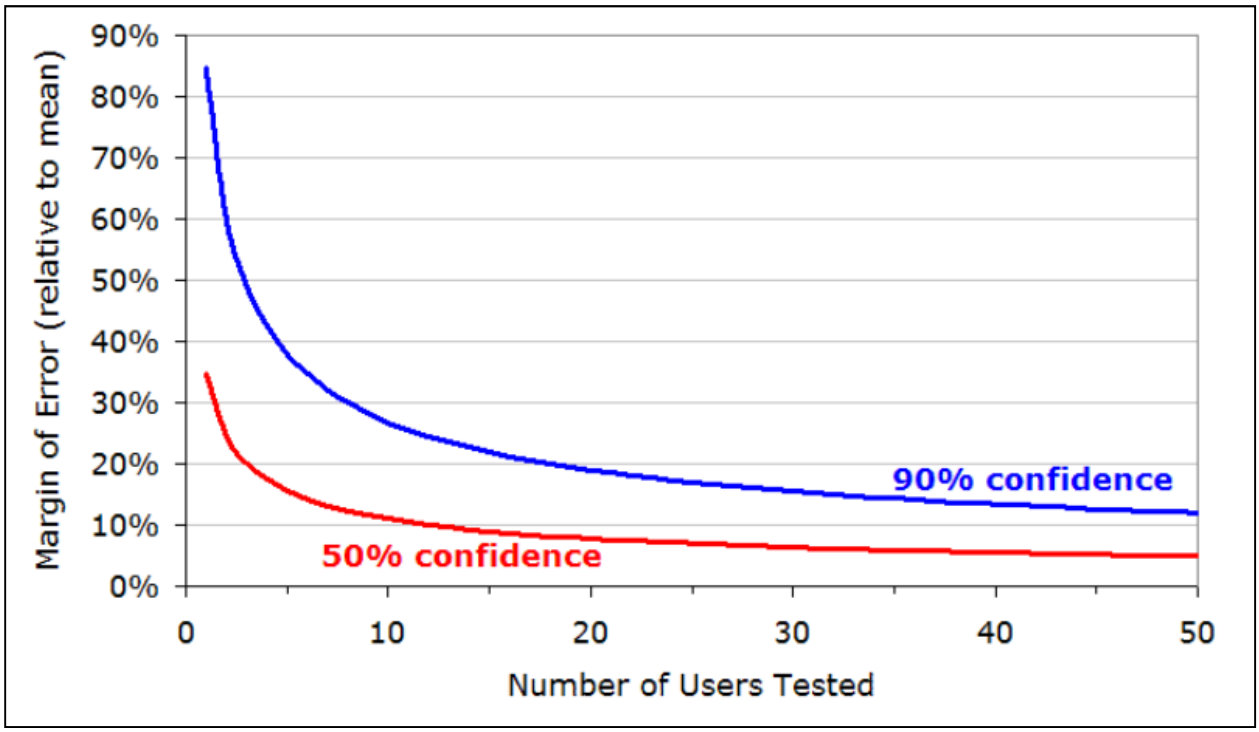

Figure 25: Margin of error for testing various numbers of users [18]

The plot represented in Figure 25 lists margin of error values on the y-axis versus the number of users tested on the x-axis. A blue curve represents the number of user's necessary for '90% confidence' in the data, while the red curve represents the number of subjects necessary for only a 50% confidence level. Thus, in order to have 90% confidence of a studies result with a +/-20% margin of error, 19 users are needed [18].

Nielsen argues that while the worst case margin of error is +/-19% for 19 users, in reality fifty percent of the time the confidence interval will be +/-8%, as shown above in figure 25. In fact, in order to halve the worst case uncertainty from +/-19% to +/-10%, a group of 76 users (71 'for data' plus the 5 outliers on average) is necessary [18]. Such a massive shift in resources is prohibitive for a number of reasons, especially for such a low reduction in uncertainty. It is unknown if similar calculations or analyses were performed in the STI or ERAU studies in order to determine sample size.

7.3 PILOT FEEDBACK – TLX, HQR

Since an ideal retest seeks to provide detailed data on the power frequency question, it is worthwhile to design such a retest while remembering lessons of the past. Specifically, when STI researchers sought to qualify their conclusions they used both Cooper-Harper scores and Pilot Induced Oscillation Rating feedback to do so. Moving forward, the use of both Cooper-Harper HQR and NASA TLX responses is critical.

When conducted electronically, the NASA TLX survey is quick to complete and is a minimal distraction to the pilot. Cooper-Harper feedback can similarly be gathered in a very short amount of time, and by gathering both data points nearly simultaneously after each run an important condition is met. While logically the premise of task loading feedback matching up

with power frequency data makes sense – especially given that handling qualities data matches – the two have never been tested in parallel when compared to power frequency. By gathering feedback in two somewhat similar disciplines of pilot opinion, such a retest will be of great value to 'bridge the gap' between task loading and handling qualities. Ideally, such a retest will show Cooper-Harper scores continue to be tied with power frequency (thus reinforcing the strength of STIs prior work) and an associated TLX trend which also matches similar expectations.

While the Bedford Scale and PIO Scale were discussed at length in section 3, their inclusion in a subsequent retest would likely do more harm than good: although handling qualities and workload may have been examined in parallel in the past, no literature was found indicating such a prior experiment. As such, a test combining the two would be the first of its kind, and adding in a third scale for comparison would have the potential to disrupt accurate data collection.

Through related research in the ICEFTD simulator, the importance of gathering verbal pilot feedback has also been reinforced. While written comments and computer-calculated scores are the basis of this research, by observation pilots tend to reveal subtle cues into their decision making processes. While such a sort of data gathering is not strictly scientific in nature, making note of pilot comments to compare them to data trends can help researchers build understanding. In addition to the aforementioned notes, it is similarly critical to develop and use a standardized briefing script for the NASA TLX and Cooper-Harper procedures.

Perhaps most importantly, the TLX data set used for this research (potentially) has a fatal flaw when examined in detail. Pilots in the 2009 study conducted a full instrument approach procedure and then provided TLX feedback. Pilots were not directed to evaluate merely the final approach segment of the test or any other segment, but instead to provide an overall score based

on their approach performance. As a result, a measure of variability was introduced into the results – for example, if pilots were instead directed to evaluate only the shorter FAF-DA portion of the approach, scores could be closely coupled to a specific region of performance.

Such direct analysis was not the desired end result of the 2009 study, but STI researchers choose to gather pilot feedback immediately after each approach with different results. In a future retest it is important to ensure that pilots are directed to provide TLX/ HQR data only within a particular region of reference (from the final approach fix to the decision altitude, for example) and that feedback is recorded in a timely manner without distraction.

7.4 'SEGMENT 5'

For simplicities sake the instrument landing task was broken down into five segments, starting with straightforward tasks of holding airspeed, altitude, and heading in segment one and terminating with an instrument approach in segment five. While segments 1-4 do yield valuable data, for the purposes of evaluating the approach in terms of establishing a correlation between power frequency and task loading it is ideal to focus on segment 5 alone.

By paring down a run from the original 12 minute long approach to a simply 'segment 5' approach, the amount of data generated decreases by approximately half. Since data from the ICEFTD was gathered at 10 Hz (STI data in the Learjet was gathered at 100 Hz), by reducing the run duration data collection speed could be increased all while generating similarly sized final data files. In addition it becomes far easier for a pilot to evaluate and recall their thoughts and actions for the TLX or HQR survey afterwards by flying an abbreviated approach.

7.5 QUALIFICATIONS

In order to "test if ICEPro had utility for mitigating a potentially hazardous icing encounter", the researchers selected a group of 29 pilots with "relatively similar flight experience". As such, all were instructor pilots holding instrument, commercial, and multi engine ratings and all had no less than 1300 hours. In addition, none had specific icing training or in-flight experience in actual icing conditions, a key facet for the ICEPro validation test.

While the icing knowledge/ experience criteria are certainly not applicable moving forward, the high standards used in pilot selection are certainly worth repeating. By identifying experienced pilots with a fairly high baseline of experience, a researcher can prevent from muddying the waters with meaningless data. While no one type of pilot is better or worse to test the correlation between power frequency and NASA TLX, the possible exception is that a higher time or more experienced pilot can provide a more accurate workload feedback assessment than his or her lower time peer.

7.6 STANDARDS

During the initial test runs, pilots were directed to attempt to sustain airline transport pilot (ATP) standards during the approach which meant maintaining airspeed within +/-5 knots and altitude within +/- 100 feet [19]. While these standards are indeed useful to drive pilot activity and to provide a desired level of performance, expanding expectations in the future is also valuable.

For example, in addition to attaining the aforementioned (ATP) standards for airspeed and altitude, adding in similar tolerances for glide slope and localizer parameters: perhaps +/-1 dot for glide slope and +/- 5 dot for localizer. Another change is to direct some pilots to attempt

to fly so-called 'ATP/2' numbers during the approach. While technically possible to achieve, the markedly increased expected level of performance would artificially force the pilot to fly aggressively in order to attain the higher standards. The additional control activity could generate higher power frequency values which in turn gives the data better fidelity.

As a compromise, an ideal retest might involve half the pilots maintaining ATP standards while the remainder strove to perform at the 'ATP/2' level. By splitting the group but holding all other factors constant the experiment sheds light on a wider scope of performance, all the while seeking to confirm the original hypothesis.

7.7 SETUP AND TRAINING

By building a training plan which directed each pilot fly a familiarization run first and then three subsequent evaluative runs for data, the ERAU researchers appropriately leveraged the element of time to their advantage. First, such a set up allowed each pilot to gain enough experience with the simulation environment to feel comfortable, yet not too much total time in the simulator that they were exhausted and performing poorly by their final sortie.

In addition, the technique allowed the researchers to efficiently evaluate a fairly large group of test subjects without gathering too little data (two runs for data per pilot) or too much meaningless information (two practice runs, three runs at ATP standards, and three runs at 'ATP/2' standards per pilot).

In subsequent NASA pilot research performed in the ICEFTD such a performance 'sweet spot' was further confirmed. Moving forward using a similar approach of several familiarization runs followed by three evaluative runs for data is recommended.

7.8 DOUBLE BLIND

During the prior experiment, the baseline and evaluative pilots were given a cursory briefing detailing the hazards of airframe and tailplane icing, the performance of the ICEFTD simulator, and for those in the 'evaluative' group, the manner in which the ICEPro software worked. None in the group were test pilots and none knew the broad-based desired or expected outcomes of the experiment. These aforementioned qualifications are critically important as they help eliminate bias and ensure that the experiments results are scientifically acceptable.

However, in a subsequent retest it would be ideal for the evaluation pilots, the researchers conducting the experiment, and even the data analysts to be blind to the desired outcome of the test. Although it adds a layer of complexity – i.e. training additional personnel to a standard where they can adequately conduct the data collection – double blind testing eliminates many potential sources of bias [20].

7.9 ICEPRO

While critical to the main premise examined during the original 2009 study, the presence or absence of the ICEPro display could potentially influence the topic at hand. After examining the available data, no improvement or decline in power frequency/ Cooper-Harper correlation seems to occur with or without the ICEPro display.

As a matter of principle a subsequent ideal retest would have all pilots flying the approach using identical control setup and displays following the one-factor-at a time (OFAT) testing model. It is reasonable to expect that lower Cooper-Harper scores and lower power frequency values would have occurred for the pilots using ICEPro since they flew using a task-focused icing stability tool at their fingertips. In contrast, the non ICEPro pilots would be

expected to exhibit the opposite trend, constantly fighting to remain in control with less feedback and recording higher workload assessment numbers due to the increased stress and diminished handling qualities.

Even this seemingly simple trend did not develop in the data, further underscoring the need for an ideal retest to examine the concept in detail. One drawback to the ICEPro system is the need for constant control inputs to update the model in real time – without those updates ICEPro cannot provide accurate performance degradation cues to the pilot. The version of the ICEPro software used in the 2009 work also used computer-controlled inputs to update the model if natural pilot inputs were insufficient, but these uncommanded control movements clearly affected calculated power frequency values.

Were the test to be conducted again in the ICEFTD it would be wise to have all pilots fly with the ICEPro displays turned 'off' and simulated icing turned 'off'. Such a configuration would provide a sterile cockpit environment, remove distractions, and refocus the experiments attention on the topic at hand.

7.10 CONTROL LOADING

While the NASA ICEFTD simulator is an incredibly useful icing research tool, one area where it does lack is control loading. Clearly an ideal flight research simulator would have control loading in all three axes in order to create a high fidelity and operationally representative environment, but when NASA directed Bihrle Applied Research to build the simulator control loading was only integrated in a single axis due to ease of construction and cost motivations.

In order to quantitatively test whether control loading would or would not measurably affect a pilots performance for this specific test, several test runs were conducted. A pilot who

had intimate knowledge of the system and the approach was selected, and two abbreviated approaches were flown. The first abbreviated approach was flow with the control loader 'on' and started just prior to the final approach fix, ending shortly after the decision altitude was reached. After simply turning the control loader system off the second approach was flown under the same simulated conditions.

		Latstick			
Pilot	**Loader**	**Avg cutoff**	**Max cutoff**	**Avg power**	**Max power**
A	On	1.22880	2.39640	0.00176	0.00583
A	Off	1.31960	2.36690	0.00241	0.01001
		Lonstick			
Pilot	**Loader**	**Avg cutoff**	**Max cutoff**	**Avg power**	**Max power**
A	On	0.57900	1.89350	0.00189	0.01092
A	Off	1.13250	3.17680	0.00429	0.00992
		Rudder			
Pilot	**Loader**	**Avg cutoff**	**Max cutoff**	**Avg power**	**Max power**
A	On	1.22260	2.39190	0.00154	0.00512
A	Off	1.30220	2.30840	0.00228	0.00954

Figure 26: Cutoff/ power frequency values for control loader runs

As is reasonable to expect, power and cutoff frequency values were elevated during the second run – several of those numbers are summarized above in figure 26. In the axis of note (LONSTK or elevator) both average and maximum cutoff frequency figures doubled or sometimes increased to much higher figures. While overall performance during the approaches was close to tolerances, the amount of energy expended by the pilot dramatically increased when they transitioned from an accurate to sloppy control loading environment.

While the important effects are evident in the elevator channel, an interesting trend also occurred tied to aileron and rudder inputs. Almost without exception, power and frequency

numbers increased in the aileron and rudder axes even though the only change from one run to another was turning the elevator control loader off. In fact maximum power frequency doubled for the ailerons when the elevator control loader was turned off, proof that the pilot's differing response in one axis translated into additional activity in all three axes.

Based on the data and calculations it is clear that a significant difference occurs when control loading is taken away while trying to gather in depth understanding of a pilots response. While a formal assessment was not conducted, pilot comments throughout both test runs strongly reinforce the aforementioned assertion. Furthermore, the original STI comparison of power frequency and Cooper Harper HQR was performed aboard an 'in-flight simulator' Learjet which clearly has highly accurate force-feel feedback in all axes of flight. It is, therefore, fair to conclude that control loading plays an important part in generating meaningful test data.

7.11 SUMMARY

1	Design of experiments	1.1	20x pilots
2	TLX, HQR	2.1	Complete NASA TLX and Cooper-Harper HQR
		2.2	Complete TLX/ HQR immediately following each approach
		2.3	Develop standardized briefing script fot TLX/ HQR
		2.4	Take notes regarding pilot survey feedback
3	"Segment 5"	3.1	Fly approach from FAF to DA only
4	Qualifications	4.1	ATP/ instrument qualification
5	Standards	5.1	Split total group into two subgroups
		5.2	'ATP' and 'ATP/2' subgroups (10x pilots each)
6	Setup and training	6.1	Each pilot flies 1-2 practice approaches, 3x for data
7	Double blind	7.1	Conduct double blind study if feasible
8	ICEPro	8.1	Conduct approach with ICEPro display/ simulated icing "OFF"
9	Control loading	9.1	Utilize control loading in all axes if able

Figure 27: Summary of changes for ideal retest

In summary, through the implementation of several key changes from the ERAU ICEPRO validation test model a new ideal retest can be performed. A number of recommendations are outlined above in figure 27.

In a truly perfect world an ideal retest would closely mimic the original STI research by using a handful of test pilots with Cooper-Harper/ TLX prior experience and use of the Calspan variable stability Learjet. Due to a number of factors including anticipated experiment cost and qualified pilot availability, the ideal retest scenario outlined presented here is likely the best available compromise.

8. CONCLUSIONS

Power frequency is a relatively new parameter which is a derivative of cutoff frequency and adds depth into the frequency and magnitude of a pilots control inputs. After much study, a positive correlation was observed between power frequency and Cooper-Harper handling qualities ratings using flight test data gathered by Systems Technology Incorporated. In order to provide flight test researchers another tool for analyzing measured pilot performance and feedback, it was important to verify whether task loading pilot feedback and power frequency values were similarly coupled.

Using data collected from eighty individual instrument approaches conducted in the NASA ICEFTD simulator, power frequency calculations were performed. Of particular interest were values for average and maximum power frequency during all phases of the approach. Given the inherent pluses and minuses of the simulator environment used for data collection, as well as the original intent of the study the data was gathered to support, special care was taken to identify factors which might influence the eventual outcome of the test.

While no clear correlation between TLX and power frequency was visible after extensive analysis, the use of a recycled data set complicated study of the fundamental issue. As a result, a list of changes for a subsequent "ideal retest" were collected and presented in detail. Although the particular data set studied did not seem to confirm a connection between task loading and power frequency as was expected, it is believed that further research can yield a clearer answer to the question.

REFERENCES

[1] "Simulation Development and Support Success Stories: Ice Contamination Effects Flight Training Device (ICEFTD)" [online database], URL: http://www.bihrle.com/services_sds_success9.html [cited 15 January 2014]

[2] "Flight Research Building Gallery" [online database], URL: http://facilities.grc.nasa.gov/hangar/gallery.html [cited 15 January 2014]

[3] Ranaudo, R., Martos, B., and Barnhart, B., "Piloted Simulation to Evaluate the Utility of a Real Time Envelope Protection System for Mitigating In-Flight Icing Hazards", AIAA 2010-7987, 2010.

[4] Lampton, A. and Klyde, D.H., "Power Frequency – A New Metric for Analyzing Pilot-in-the-Loop Flying Tasks", AIAA 2011-6539

[5] "Aviation Troubleshooting: Future of Deicing Technology and Effective Training for Flight in Icing Conditions" [online database], URL: http://aviationtroubleshooting.blogspot.com/2011_01_01_archive.html [cited 15 January 2014]

[6] "NASA TLX Homepage" [online database], URL: http://humansystems.arc.nasa.gov/groups/tlx/ [cited 17 January 2014]

[7] "NASA TLX Homepage: Paper/ Pencil Version" [online database], URL: http://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf [cited 17 January 2014]

[8] "NASA TASK LOADING INDEX (TLX) v. 1.0" [online database], URL: https://wiki.cc.gatech.edu/ccg/_media/classes/muc/tlx_manual.pdf?id=classes%3Amuc%3Afall08%3Areadings&cache=cache [cited 17 January 2014]

[9] Cooper, G.E., and Harper, R.P., "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities," Advisory Group for Aerospace Research and Development, 1969.

[10] "Figure 66: Cooper-Harper Handling Qualities Rating Scale" [online database], URL: http://history.nasa.gov/SP-3300/fig66.htm [cited 1 February 2014]

[11] Harris, D., Gautrey, J., Payne, K., and Bailey, R., "The Cranfield Aircraft Handling Qualities Rating Scale: A Multidimensional Approach to the Assessment of Aircraft Handling Qualities", Royal Aeronautical Society, 1968.

[12] "Modified Cooper-Harper Scales for Assessing Unmanned Aerial Vehicle Displays" [online database], URL: http://web.mit.edu/aeroastro/labs/halab/papers/MCHUVD.pdf [cited 1 February 2014]

[13] Roscoe, A.H., and Ellis, G.A., "A Subjective Rating Scale for Assessing Pilot Workload in Flight: A Decade of Practical Use", Royal Aerospace Establishment, 1990.

[14] "Annex C – BEDFORD WORKLOAD SCALE" [online database], URL: http://ftp.rta.nato.int/public/PubFullText/RTO/AG/RTO-AG-300-V27/AG-300-V27-ANN-C.pdf [cited 1 February 2014]

[15] Roscoe, A.H., "HEART RATE AS AN IN-FLIGHT MEASURE OF PILOT WORKLOAD", Royal Aerospace Establishment, 1982.

[16] "MIL-HDBK-1797: FLYING QUALITIES OF PILOTED AIRCRAFT", Department of Defense, 1997.

[17] "FTM-107: U.S. NAVY TEST PILOT SCHOOL FLIGHT TEST MANUAL, ROTARY WING STABILITY AND CONTROL", Department of Defense, 1995.

[18] Nielsen, J., "Quantitative Studies: How Many Users to Test?" [online database], URL: http://www.nngroup.com/articles/quantitative-studies-how-many-users/ [cited 25 March 2014]

[19] "Airline Transport Pilot and Aircraft Type Rating – Practical Test Standards for Airplane" [online database], URL: https://www.faa.gov/training_testing/testing/test_standards/media/FAA-S-8081-5F.pdf [cited 12 February 2014]

[20] Kosmulski, M., "Skeptical Comment About Double-Blind Trials", *The Journal of Alternative and Complementary Medicine*, 2010.

APPENDIX

APPENDIX A
Power frequency plots, segments 1-5

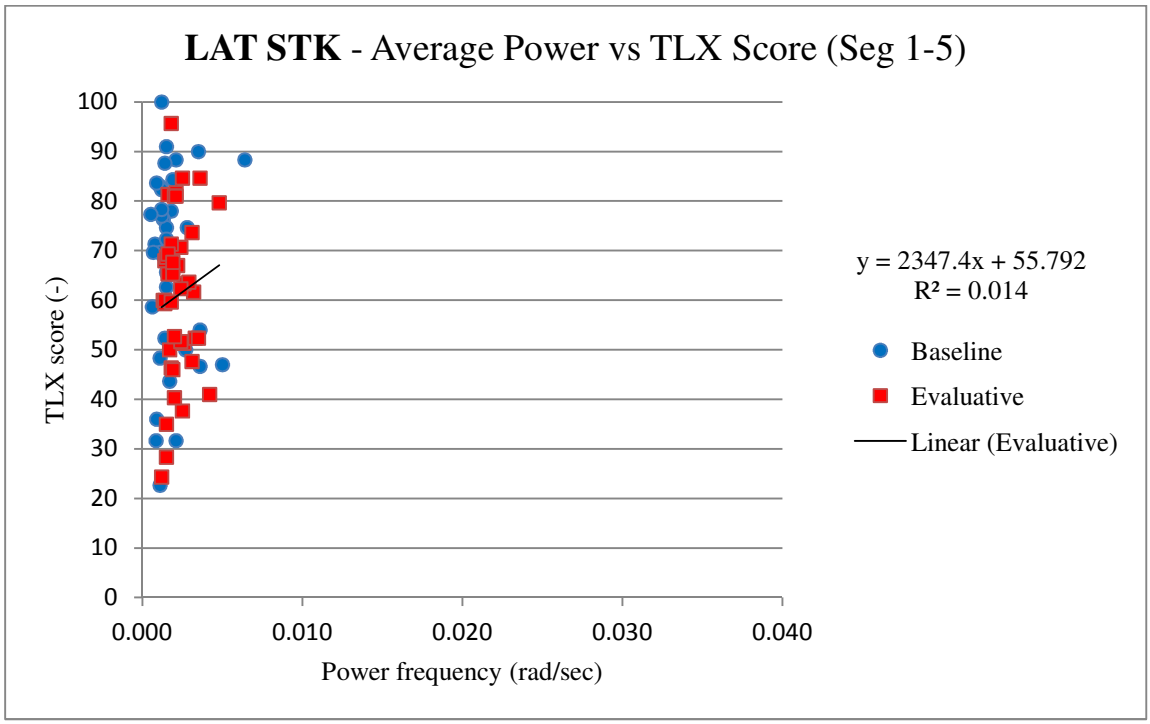

Figure A.1 – Aileron, average power vs TLX (Seg 1-5)

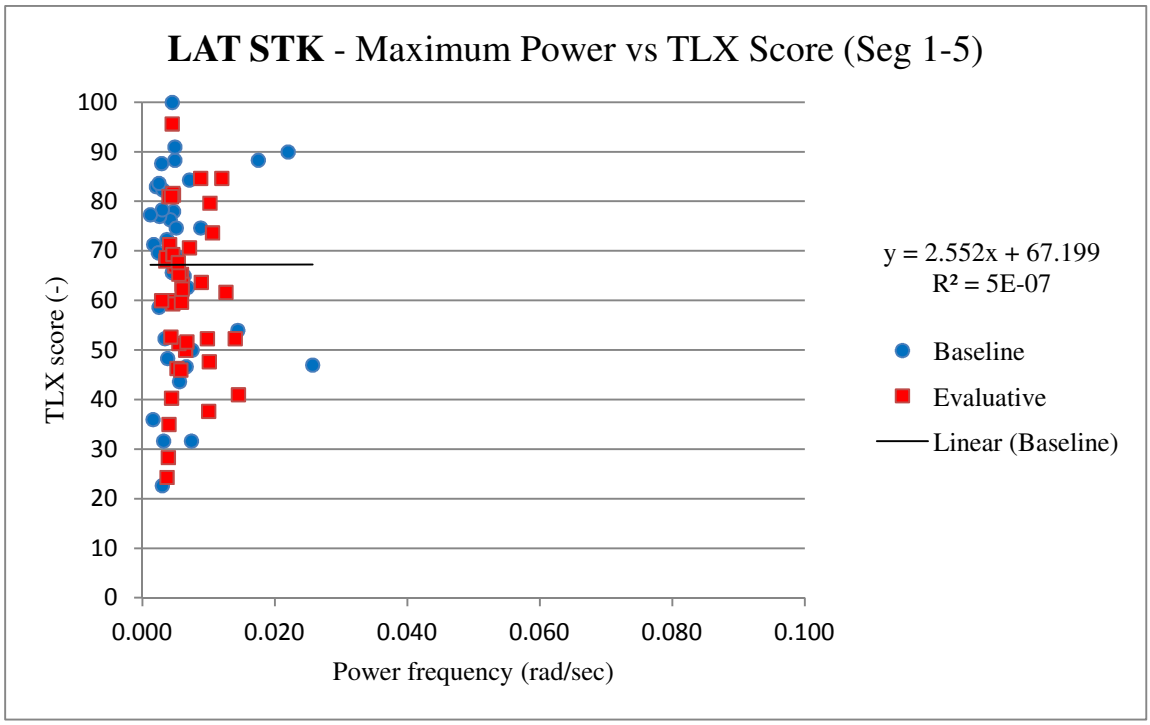

Figure A.2 – Aileron, maximum power vs TLX (Seg 1-5)

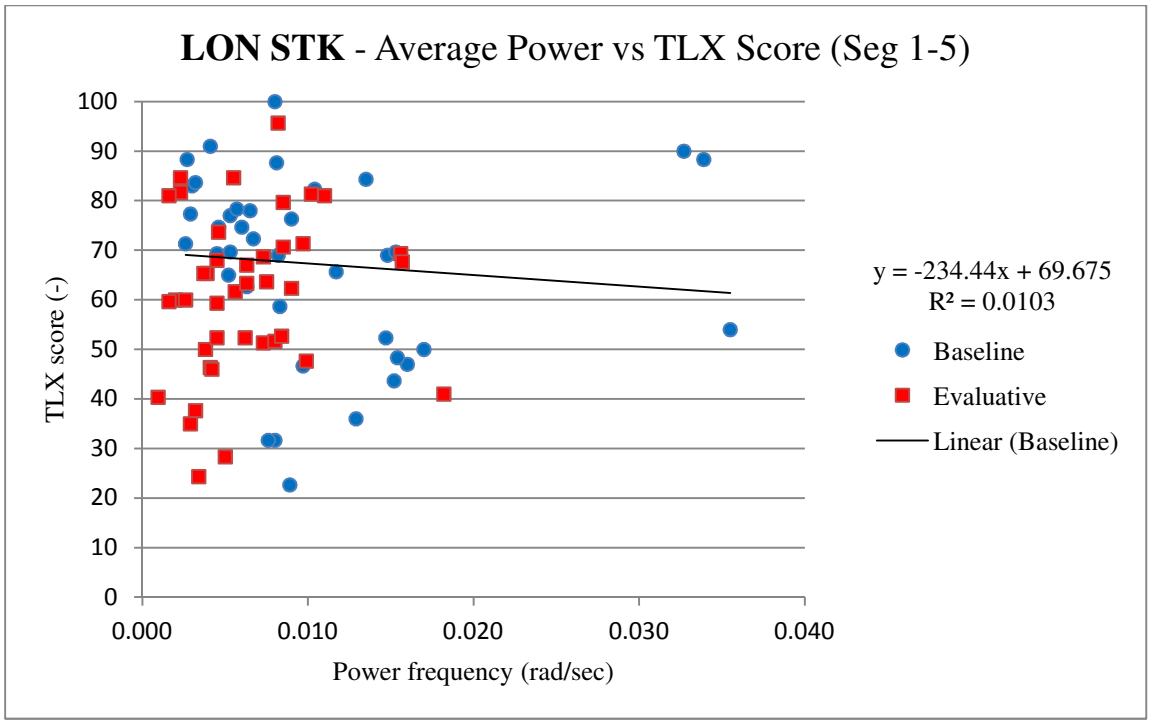

Figure A.3 – Elevator, average power vs TLX (Seg 1-5)

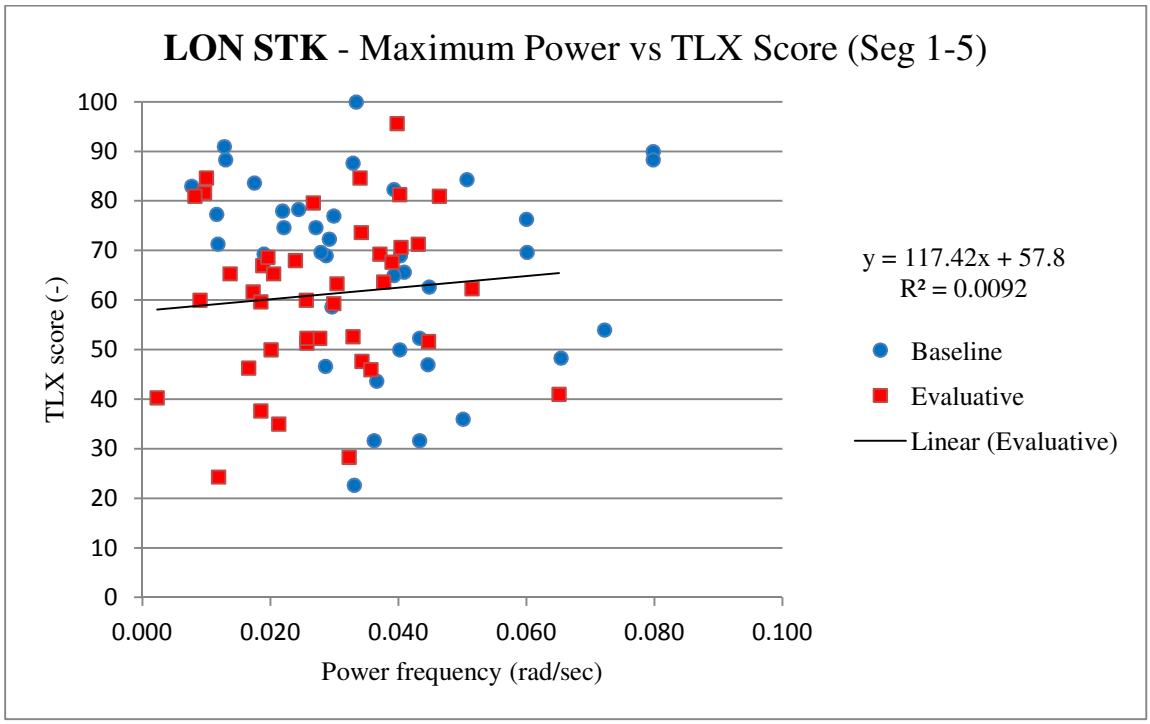

Figure A.4 – Elevator, maximum power vs TLX (Seg 1-5)

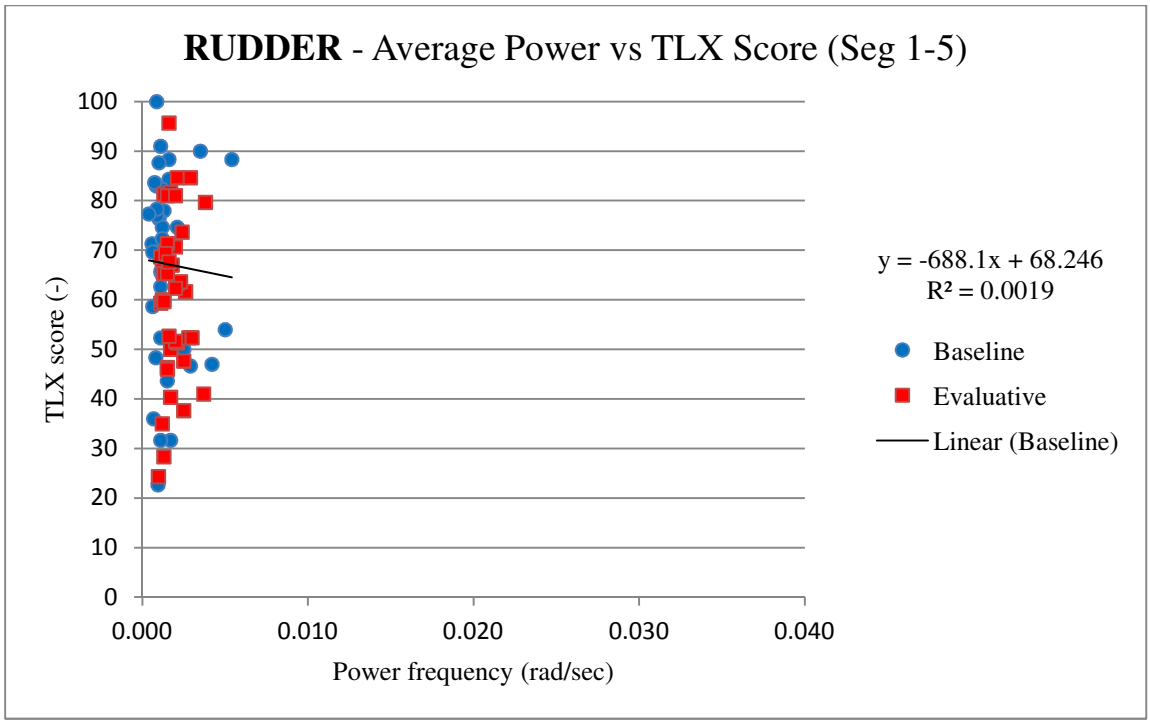

Figure A.5 – Rudder, average power vs TLX (Seg 1-5)

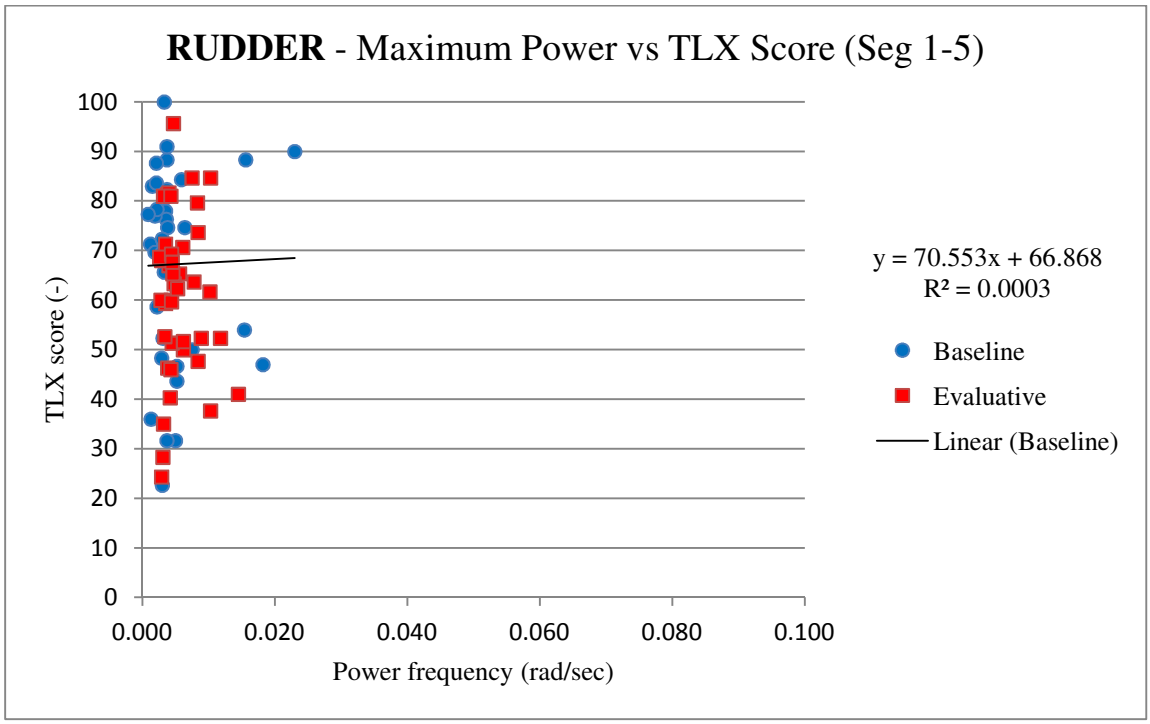

Figure A.6 – Rudder, maximum power vs TLX (Seg 1-5)

VITA

Antonio Gemma Moré grew up in Tullahoma, TN, the son of Marcos Ortiz Moré and Kathleen Brenda Gemma. He graduated from Tullahoma High School in 2008, earning his powered Private Pilots License in 2007 and glider rating in 2009. He then attended Tennessee Technological University in Cookeville, graduating with a Bachelors of Science degree in Mechanical Engineering in 2012. In 2012, he began the pursuit of a Master's of Science degree in Engineering Science at the University of Tennessee Space Institute in Tullahoma, TN. During his time at the University of Tennessee Space Institute, he worked as a Graduate Research Assistant performing work that included a pilot survey using NASAs Ice Contamination Effects Flight Training Device (ICEFTD). Antonio is currently pursuing his Master of Science degree in Engineering Science with a concentration in Flight Test Engineering.