Investigation of Pilot Inceptor Metrics as Correlates to Qualitative Workload

Dr. Ryan Paul
Assistant Professor
Oklahoma State University
Stillwater, OK, USA

Matthew Rhinehart
Aerospace Engineer
NAVAIR
Patuxent River, MD, USA

Kenneth Cahill
Aerospace Engineer
NAVAIR
Patuxent River, MD, USA

Zachary F. Smith
Aerospace Engineer
NAVAIR
Patuxent River, MD, USA

ABSTRACT

Advances in control system and augmentation technologies in the recent decades have allowed major advances in both the safety and mission effectiveness of military aircraft. These technologies have also resulted in an increase in the complexity of the development and certification process, partially due to the myriad of new "knobs" that can be turned. At the same time, the primary means of Handling Qualities evaluation remains subjective, via pilot assigned ratings, unchanged since Cooper & Harper published their scale in 1967. This effort sought to collect a large quantity of fundamental piloted simulation data, primarily to establish if correlation exists between qualitative rating scales and previously developed inceptor workload metrics. Secondarily, different types of classical Handling Qualities degraders were used to determine if different inceptor metrics were better able to identify them. Results indicate that most pilot subjects tested show at least some correlation between qualitative ratings and quantitative inceptor metrics, however few showed strong enough correlation to consider usurping the role of qualitative scales. Further, it showed that to achieve any meaningful correlation, treating pilots as individuals is necessary, but not sufficient, as a few pilots demonstrated poor or no correlation.

INTRODUCTION

Advances from near-peer adversaries and the drive to ever-improve warfighter safety serve to apply pressure to increase the capability of military aircraft. One major advance to answer this call has been the development of cutting-edge fly-by-wire control systems. These systems deliver unheralded flexibility and capability; however, it has resulted in significant increases in the time required to optimize the handling qualities. This is partly due to the bewildering number of possible tunable parameters that can be applied across ever-broadening flight envelopes. At the same time, the drive to develop and field aircraft faster and more cheaply are in direct opposition to the desire for additional capability. Accurately discerning the workload required by a pilot for completing a given mission task or mission task element is important for aircraft certification, development of good handling qualities, and assuring mission effectiveness. A robust and objective workload metric would not only offer quick verification of handling qualities, it could also serve to guide the control designer to which aspects of the design to tune in the development of capability, thus increasing efficiency.

Historically, pilot workload is given based on qualitative scales, such as the Cooper-Harper Handling Qualities Rating (CHR) scale (Ref. 1). The CHR Scale is a 10-point ordinal scale, and through a series of questions, guides the pilot to a rating. The CHR Scale is a composite of task performance and self-assessed workload. While task performance can be objectively measured, self-assessed pilot workload is subjective. Measurement errors stem from factors such as inter- and intra-pilot variation, non-linearities in the CHR Scale, and other issues as discussed further in Ref. 2 and Ref. 3. Other intangible challenges arise when trying to measure workload objectively such as learning effects or piloting strategy changes. In addition, differences between simulation and flight, such as modeling issues, cueing limitations, environmental factors, adherence to aircraft limitations, and self-preservation, pose an increased challenge to objectively

Presented at the VFS Handling Qualities Technical Meeting, Huntsville, Alabama, USA Feb. 7-8 2023.

measuring workload in pilot-in-the-loop simulators; which are often used during the aircraft development process. These subjective variations often limit the utility of the CHR Scale for the desired purpose, leading to the question of "how to measure workload objectively?".

Objective measurement of workload has long been of interest both in the field of handling qualities and other disciplines producing concepts spanning physiological, neurological, aircraft state, inceptor activities, and other measures as shown in Refs. 3, 4, and 5. This study continues the investigation of using pilot inceptor workload metrics as an analog to subjective workload due to ease of collection from both flight and simulation. Many inceptor measures have been proposed and evaluated to varying degrees of rigor, however none have found their way into regular use by the handling qualities community. While the general inertia of the status quo may partially be to blame, most of the proposed inceptor activity measures are not suitable for broad application, and therefore are not widely adopted. With the exception of the aggressiveness metric in Ref. 5, most metrics focus on correlation of a single axis with little discussion on application in multi-axis tasks. In addition, limited information is available in the literature regarding application of metrics to time varying tasks. Ref. 6 applies the inceptor based workload metrics that could be found (Ref. 4 and 5, and Ref. 7 through Ref. 12) to a given set of data in an attempt to determine which ones offer the greatest advantage. While initial findings were inconclusive, it was thought this was due to the task against which the metrics were run and the size of the candidate pool among other reasons. Therefore, data collected specifically for analysis against these metrics was needed.

For a relatively complicated task such as the Aeronautical Design Standard 33 (ADS-33) Precision Hover Mission Task Element (MTE) (Ref. 13), a viable metric must objectively determine workload during the 45-degree inbound translation, hover capture, and position hold phases. It is not immediately clear if the composite handling quality rating (HQR) for the task should be based on the average or peak value. The United States Naval Test Pilot School (USNTPS) does not explicitly instruct students to consider peak or average workload over the total task. This leaves an open question, which could contribute to both inter- and intra- pilot variability. Finally, a functional workload metric must be able to determine workload change stemming from different factors such as control tuning parameters, environmental effects, and cueing environment. A relevant workload metric should be scalable to any number of axes, not require knowledge of the explicitly observed error by the pilot (pursuit vs. compensatory), account for the time varying nature of some tasks, and adequately capture different drivers of workload. The study laid out within this report uses data collected on a relatively large candidate pool that had pilots fly tasks that were derived from selected MTEs with known degraders to drive up workload. Collecting and analyzing a large pool of data is expected to reveal the capability of the PIW metrics to stand-in as an analog to workload.

METHODOLOGY

In an effort to modernize the approach to measure pilot workload, this research followed a novel path for analyzing pilot inceptor data to produce results that correlate with perceived workload. This work was predicated on the hypothesis that either a singular metric or population of metrics exist that can be computed from recorded data whose values correlate with qualitative pilot workload. Following an initial examination using historical data from a previous research endeavor (Ref. 6), the correlation between qualitative workload and computed metrics was weaker than desired and attributed in part to being restricted to only HQRs as the measure of qualitative workload. The project then shifted focus and emphasized developing and executing a clinical set of tasks in a simulator to build a more comprehensive database of qualitative pilot assessments using three different techniques for capturing pilot workload. The final product was a database of 12 subject matter expert pilots performing up to 5 tasks in a pilot-in-the-loop simulator to provide more insight into the relationship between the task, inceptor motions, and perceived workload. Each MTE was performed multiple times with varying aircraft configurations to elicit higher workload in manners known to degrade aircraft handling characteristics as documented in ADS-33. The research goal was to assess the appropriateness of various pilot inceptor workload (PIW) metrics and learn how to apply them in a test environment.

The simulation tasks flown were primarily designed to collect a broad set of data to conduct research into an objective measurement of pilot workload. The five tasks selected emphasized differences between single vs. multiple axis tasks, compensatory vs. pursuit tasks, pseudo-time variant vs. time varying tasks, and fine vs. large control input tasks. Each task was first flown using a nominal air vehicle configuration, and subsequently repeated, each time adjusting the underlying model to exhibit classical degraders of handling qualities, per ADS-33 specifications.

The Simulation Environment

The host lab for this experiment was the single seat fixed base simulator at the USNTPS. The cockpit was generic and equipped with inceptors capable of supporting fixed or rotary wing simulations. A control loader provided functionality to adjust mechanical characteristics of the inceptors. The typical center stick and collective helicopter controls were used for this evaluation. A flat panel computer monitor in the cockpit provided representational flight instruments, while a large flat projection, approximately 8 feet square, presented the pilot with an out the window visual scene generated with a single ceiling mounted projector. Figure 1 shows the cockpit arrangement and Figure 2 exhibits the typical setup during data collection.

Figure 1. Simulator Pilot Seat, Controls, and Primary Flight Display.

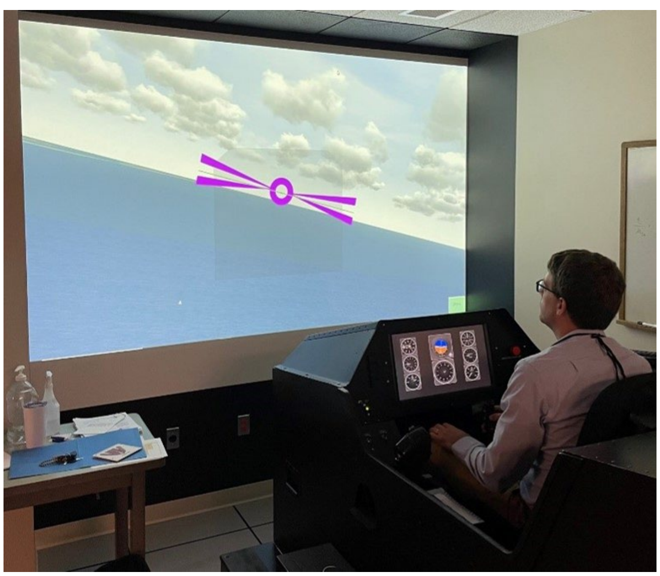

Figure 2: Overall Simulation Configuration.

The flight dynamics model for this research was of state space form and loosely based on the OH-6. Several underlying assumptions limiting the applicability of the model are documented here. All of these assumptions can be removed by increasing the complexity of the model, but at the cost of loss of insight and ease of customization. The significant assumptions for the model are as follows:

- The model is for a fixed operating point with constant coefficients. Stability derivatives are used to define the response characteristics.
- The model is valid only for hover and low-speed flight.
- Some stability derivatives have been neglected. The most significant of these are the static stability derivatives Mw and Nv, along with the speed-dependent derivatives Zu, Mu, and Lv. The main consequence of this simplification is that the free response dynamics are not in the form of the familiar hovering cubic that is typical for three degree of freedom models (Ref. 1).
- All inter-axis coupling derivatives are set to zero. Placeholders are identified for pitch-to-roll and roll-to-pitch coupling due to aerodynamics (Mp and Lq) or control mixing (M_A and L_B). These derivatives should be set to achieve the desired amount and type of coupling desired, as discussed later in this paper.
- Control derivatives are considered variables to be set in simulation to provide best response for each set of dynamics. Alternatively, multipliers can be included in the simulation environment allowing for adjustment of control sensitivity without modifying the values of the derivatives.

A first-order state-space representation of a hovering helicopter with fixed aerodynamic parameters comprises the basic math model. Derivatives that define oscillatory low-frequency (phugoid) dynamics are not included, though it is possible to increase model complexity and add those derivatives. Capability for pitch-due-to-roll and roll-due-to-pitch coupling, resulting from both dynamics and control coupling, is included in the model. In addition, an Attitude Command model was generated by adding angular attitude derivatives to the equations of motion. The resulting generalized state-space matrices for an $\dot{x} = Ax + Bu$ state transition form are as given below, included as Figure 3:

$$A = \begin{bmatrix} X_u & 0 & 0 & -g & 0 & 0 & 0 & 0 \\ 0 & Z_w & U_0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & M_q & M_\theta & 0 & M_p & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & Y_v & 0 & -U_0 & g \\ 0 & 0 & L_q & 0 & 0 & L_p & 0 & L_\phi \\ 0 & 0 & 0 & 0 & 0 & 0 & N_r & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & Z_{\delta_C} & 0 & 0 \\ M_{\delta_B} & 0 & M_{\delta_A} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ L_{\delta_B} & 0 & L_{\delta_A} & 0 \\ 0 & 0 & 0 & N_{\delta_P} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$X = \begin{bmatrix} u \\ w \\ q \\ \theta \\ v \\ p \\ r \\ \phi \end{bmatrix} \quad \text{and} \quad u = \begin{bmatrix} \delta_B \\ \delta_C \\ \delta_A \\ \delta_P \end{bmatrix}$$

Figure 3: State Space Matrices.

Variations in vehicle response (bandwidth) are made by changing the values of the dynamic derivatives. Control derivatives in the B matrix were adjusted in simulation to provide best control/response based on pilot comments during model development. Simplifications to the baseline A and B matrices were made to establish a canonical model with fully decoupled state and control matrices. All modal dynamics were along the real axis in an s-plane representation. Artifacts such as rotor delay and higher order lag states were excluded, however an unstable long period pitch and roll mode was incorporated to reflect the natural hovering instability of a helicopter. Implementation of the unstable long-period dynamics in pitch and roll was done by introducing Mu and Lv terms to the A matrix.

As a baseline, the resulting model was tuned to fall near/on the level 1 boundary of ADS-33 criteria, as shown for short-term response criteria and pitch/roll oscillation limits in Figure 4. To accommodate analytical handling qualities bandwidth determination, a delay term was added to the baseline model to drive the phase curve to roll off beyond -180 degrees at high frequencies. The delay value selected was 1/60 seconds, which was the approximate update rate of the visual projector. No delays were added to the Simulink model, which executes in real-time during testing, as the equivalent visual delay was present while running the experiment. There was no ambient wind present in the simulator environment.

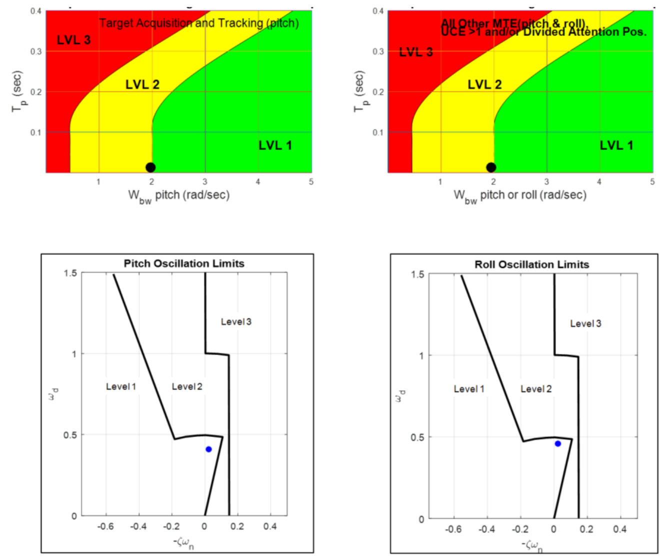

Figure 4: Baseline Model Characteristics.

To drive workload, the baseline Level-1 model was degraded consistent with deficiencies documented in ADS-33. Modifications to the plant model stability, short term response bandwidth, cross coupling, and response type were enacted simply by modifying terms in the A and B matrices defined previously. The nonlinear model modifications, an introduction of phase delay and a rate limit, were enacted directly in the underlying Simulink model. Table 1 summarizes the various modifications applied to the model and explored during the testing.

Table 1: Model Modifications.

Modification	Description
Reduced Stability	Degraded overall model to level 2 HQ
Phase Delay	Introduced 0.2s time delay
Response BW	Degraded BW to level 2 HQ
Cross Coupling	Increased coupling to level 2 HQ
Rate Limit	Introduced a 1”/s rate limit on the cyclic
Stick Force	Increased stick force gradient to level 1 limit
Response Type	Enable attitude response type

Task Descriptions: Sum of Sines - Pitch, Roll, Pitch& Roll

The Sum of Sines (SoS) task was driven by an automated command signal generated by the simulation environment. From steady, wings level flight, pilots were instructed to aggressively track the displayed signal and attempt to keep errors within the specified tolerances specified in Table 2, as adopted from Ref. 13. This maneuver did not require a test course but rather a visual cue with the desired and adequate performance criteria displayed to the pilot to enable real-time compensatory tracking. The length for scoring time was 60 seconds. There was a 10 second ramp in/ramp out period on each end of the 60 second scoring window. Figure 5 is representative of the cues provided during the sum of sines task. The pitch indication is the green dot at the center of the image. Desired for pitch is to keep the green dot within the inner circle while adequate is to keep it within the magenta circle. The roll indication is the green horizontal line. Desired for roll is to keep the green line within the inner most wedge shape while adequate is to keep it within the magenta wedge shape. This task was driven by an automated command signal generated by the simulation environment and followed a randomized sum of sines command.

Table 2: Desired and Adequate Tolerances.

	Desired	Adequate
Pitch: At least X% of the scoring time	50%	75%
within pitch attitude error tolerance:	±1°	±2°
Roll: At least X% of the scoring time	50%	75%
within roll attitude error tolerance:	±5°	±10°
PIO Considerations	No PIO tendencies	No divergent PIO tendencies
Inter-axis coupling shall not be:	Undesirable	Objectionable

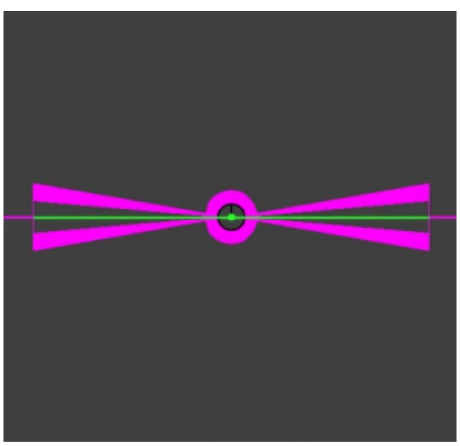

Figure 5: Sum of Sines Cueing.

Task Descriptions: Point to Point Reposition

Pilots were initialized in a 20-foot hover over the runway threshold on centerline and holding the runway cardinal heading. They were then instructed to maneuver the aircraft to arrive in a hover at an altitude of approximately 20 feet, 1000 feet down the runway on a reciprocal heading and on centerline. At some point during the maneuver, the pilot was to climb to approximately 100 feet. The pilot would call "mark" when they departed the initial hover location and "mark" again when they arrived at the second hover location to the pilot's satisfaction. The test course utilized standard ICAO runway markings as shown in Figure 6. As this task was intended to allow the pilot flexibility in how it is completed (and to what accuracy), no absolute performance standards were prescribed. The description of the maneuver specified beginning and ending positions and headings, as well as an intermediate altitude, but the pilot was asked to follow these directions to their own level of acceptable tolerances. An order of operations was not specified (e.g. the pilot can choose to effect the heading change or climb to 100 feet at any time during the maneuver). The elapsed time from "mark" to "mark" was, recorded however no time requirements were dictated and there should be no assumption that a faster time was better. The pilot was expected to use their nominal control strategy and let the natural level of workload fallout.

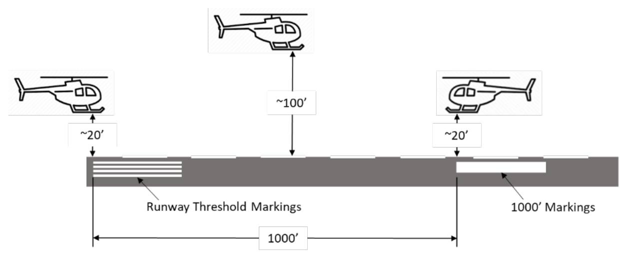

Figure 6: Runway Markings.

Task Descriptions: Precision Hover

The pilot initiated the maneuver at a ground speed of between 6 and 10 knots, at an altitude less than 20 feet. The target hover point was oriented approximately 45 degrees relative to the heading of the rotorcraft. The target hover point was a repeatable, ground-referenced point from which position deviations were assessed. The test course for this maneuver is depicted in Figure 7. Note that differences between the eye-point longitudinal position relative to the target longitudinal position will effectively change the tightness of the lateral and vertical tolerances (closer equals larger tolerances). Pilots accomplished the transition to hover in one smooth maneuver. It was not acceptable to accomplish most of the deceleration well before the hover point and then to creep up to the final position.

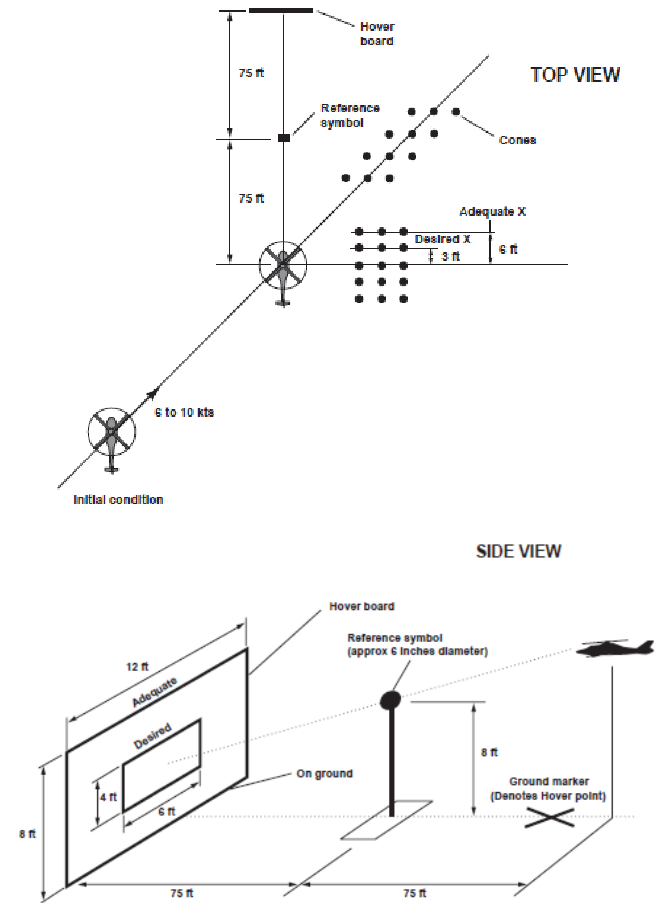

Figure 7: Precision Hover Course.

Subject Matter Expert Characterization

The test group was a diverse body of pilots from civilian and military training. Civilian pilots were all at least private pilot rated, three of which held instrument ratings with one holding a commercial flight instructor rating. Two of the civilians have received formal training from a military test pilot school. All of the pilots in the civilian group primarily had single engine piston time with an average flight experience of approximately 650 hours. Military pilots were predominately recent graduates of USNTPS in their first test tour with fleet experience in fixed and rotary wing aircraft that spanned F-18E/F/G, AV-8B, P-3C, P-8A, E-2D, and UH-1Y. Exceptions include one subject who was a retired Marine Corps CH-53E fleet aviator yet another was a current UH-60L Army National Guard pilot. The average flight experience among military pilots was approximately 1600 flight hours. The test group is outlined in Table 3.

Table 3: Pilot Backgrounds.

Evaluator	Type	Primary Platform	Total Time (hrs.)	Active Pilot	Highest Aviation Credential
1001	GA	Van's RV-4	473	Yes	PPL + IR
1002	GA	PA-28	305	Yes	PPL
1003	GA/Mil	C-172/ M-20C/ UH-60L	550	Yes	Fleet Army Co-pilot
1004	Mil	P-3/P-8	1180	Yes	Test Pilot
1005	Mil	E-2	2105	Yes	Test Pilot
1006	Mil	UH-1Y	1660	Yes	Test Pilot
1007	Mil	F-18	1000	Yes	Test Pilot
1008	Mil	CH-53E	1900	No	Fleet Marine Aircraft Commander
1009	GA	C-150	1080	Yes	PPL+IR
1010	GA	C-172	60	No	PPL
1011	Mil	AV-8B	1805	Yes	Test Pilot
1012	Mil	P-3/P-8	1652	Yes	Test Pilot

Workload Rating Scales

Classically, handling qualities and more specifically workload, have been evaluated in qualitative ways. Rating scales such as Cooper-Harper, Bedford, and NASA task load index (TLX) can be used to guide pilots to a subjective rating to characterize the goodness of the system to accomplish a task. For this evaluation, pilots were asked to provide qualitative responses in accordance with selected rating scales as outlined below. Pilots were not asked to provide Bedford workload ratings (BWR) because it requires the definition of a secondary task with which to judge spare capacity. There was a concern that the addition of a secondary task could color the results by fundamentally modifying the primary task and inadvertently correlate ratings with the Plain Old Workload Scale (POWS). In other words, while providing a secondary task as a measure of primary task workload is certainly a valid approach (as with Bedford), there was concern that the secondary task would inappropriately affect the quantitative measures of workload by altering the pilots' inceptor activity characteristics. For each task and model modifier type, the pilot was asked to provide qualitative workload according to three different rating schemes: POWS, NASA TLX, and CHR.

Workload Rating Scales: Plain Old Workload Scale (POWS)

POWS, which was created for this effort, was a contrived 10-point linear scale with terminal points anchored by zero effort on the low side and maximum effort on the high side. Task performance was not explicitly considered in this rating although it was expected that the subject intrinsically considered their performance in the selection of their POWS rating. Subjective handling qualities or workload scales try to address the inherent issues in qualitatively assigned ratings by applying some form of structure to anchor results from different tests and pilots. For this effort, the possible benefit of level setting all subjects is acknowledged but its necessity is not pre-supposed. Removing the requirement that pilots' ratings should be directly comparable allows the application of a minimally structured, ordinal scale. In the case of POWS, the pilot was asked to rate their workload on a scale of 1 (being negligible workload) to 10 (cannot work any harder). It is left up to the pilot to determine the workload make-up from various sources such as physical, mental, etc. but it does not matter as long as the pilot provided a number between 1 and 10. The pilot may give half ratings, such at 4.5, if deemed necessary.

Workload Rating Scales: NASA Task-Load Index (TLX)

TLX is a NASA-developed approach (described in Ref. [1]) that first weighs a series of factors according to the pilot's perception of importance as they related to accomplishing the task. The pilot was then asked to provide a subjective rating for each factor by placing an index along a non-ordinal continuum. The TLX scale is a subjective workload analysis tool developed for generic human performance tasks. It breaks workload down into six factors (mental demand, physical demand, temporal demand, performance, effort, and frustration). A numeric score is assigned by the subject for each factor. The factors are rated in order of relative importance to allow weighing of the sub-scale scores to generate a single numeric score (with higher being increased workload). The weighting of the factors is mechanized as a set of 15 "flash cards" where the pilot selected which of the two factors shown are more important. The pilot asked to generate a weighting for each of the tasks flown and assign workload ratings for each test point flown. The factor weight and factor value was then algorithmically combined to report a score given by a numerical value between 0 and 100. Considerations to be made by the subject in each Factor area:

- Mental Demand (high/low) 1. How much mental and perceptual activity was required? (e.g. Thinking, deciding, calculating, remembering, looking, searching, etc.) 2. Was the task easy or demanding, simple or complex, forgiving or exacting?
- Physical Demand (high/low) 1. How much physical activity was required? (e.g. Pushing, pulling, turning, controlling, activating) 2. Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
- Temporal Demand (high/low) 1. How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? 2. Was the pace slow and leisurely or rapid and frantic?
- Performance (Good/Poor) 1. How successful do you think you were in accomplishing the goals of the task by the experimenter (or yourself)? 2. How satisfied were you with your performance in accomplishing these goals?
- Effort (High/Low) 1. How hard did you have to work (mentally and physically) to accomplish your level of performance?
- Frustration level (High/Low) 1. How insecure, discouraged, irritated, stressed, and annoyed vs secure, gratified, content, relaxed, and complacent did you feel during the task?

Workload Rating Scales: Cooper-Harper (CHR)

The traditional Cooper-Harper rating scale was the third and final assessment of pilot workload and served as a common baseline, both to link the new data to the original study (Ref. 6) and relate the POWS and TLX ratings to a familiar reference. The CHR scale is the classical measure by which handling qualities tasks are evaluated. The scale is bi-modal, in that it considers aspects of performance as well as workload. For tasks where desired and adequate tolerances are clearly defined, the pilot was asked to evaluate the test point in accordance with the CHR scale, shown below as Figure 8.

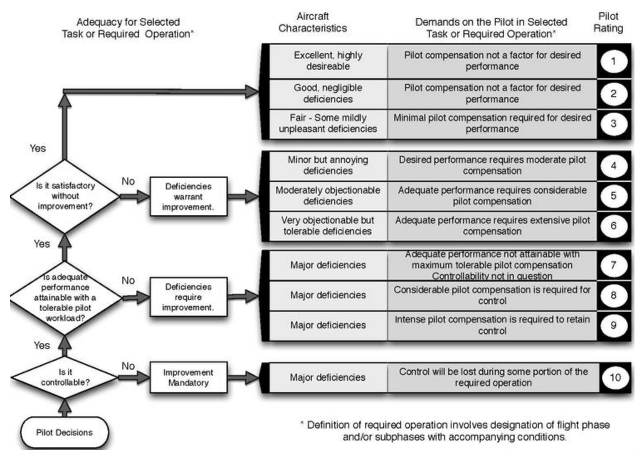

Figure 8: CHR Scale.

Data Collection

Pilots were given an introductory brief on the study's purpose, the general execution strategy for the evaluations, and the rating scales in use. At the start of each task, they were permitted to perform practice runs to gain familiarization prior to conducting formal evaluations. Additionally, the baseline model configuration was always the first presentation. Subsequent evaluations were of the various model permutations for a given task and selected at the test director's discretion. The pilot was never informed as to what particular model version they were evaluating. Following the conclusion of each trial, the recoded inceptor data was saved for post processing and the pilot was asked to provide their ratings using the three prescribed scales. The rating order of POWS, TLX, and HQR was intentionally selected as a best practice to avoid cross contamination of pilot opinion between POWS and HQR, both of which are ordinal, 10-point scales but have unique guidelines for their use.

PIW Metrics

All metrics identified in Ref. 6 were computed for the data. Two of the most promising from the earlier work, aggressiveness and mean power frequency, are highlighted herein when the results are presented. The definitions of the metrics are included for completeness.

The aggressiveness metric (from Ref. 5), denoted J in Equation 1, takes input time histories (from time t_0 to t_f) from the pilot lateral and longitudinal inceptor positions. The terms δ_{ap} and δ_{ep} represent the lateral and longitudinal stick inputs after filtering out all content above 2Hz to only retain conscious pilot intent. The terms vectors $\overline{\delta}_{ap}$ and $\overline{\delta}_{ep}$ are generated by low-pass filtering the raw cyclic displacement vectors at 0.25Hz to generate a representation of long-term control activity such as trim changes. The differencing between the inputs representing pilot intent and the long-term control activity ensures that the metric only inflates when pilot inputs are actively regulating the vehicle state in response to disturbances or actively tracking some target. Versions of the aggressiveness metric as presented below and with collective and pedal inputs added were computed. Due to the minimal amount of activity on the collective and pedal for the sum-of-sines MTEs performed, the inclusion or exclusion of these channels had almost no impact on the evaluated value of the metric for those tasks.

$$J = \frac{100}{t_f - t_0} \sum_{\tau=t_0}^{t_f} \left(\frac{\left|\delta_{ap}(\tau) - \overline{\delta}_{ap}(\tau)\right|}{\delta_{ap}^{max} - \delta_{ap}^{min}} + \frac{\left|\delta_{ep}(\tau) - \overline{\delta}_{ep}(\tau)\right|}{\delta_{ep}^{max} - \delta_{ep}^{min}} \right) \Delta\tau$$

Equation 1: Aggressiveness Metric.

Work performed by Lampton and Klyde in Ref. 4 suggested the use of wavelet transforms to facilitate the computation of cutoff frequency throughout the time history of a maneuver, and proposed a metric deemed power frequency. The coefficients produced from the wavelet transform allow calculation of the power across both frequency and time. Analogous to the calculation of cutoff frequency, the time varying cutoff frequency (ω_c (t)) is found by determining the frequency at which half the total power has accumulated for each time instance in the time-frequency representation. From Ref. 4, the time varying cutoff frequency is computed as shown in Equation 2. In this expression, the PSD curve at each time instance, $G_{\delta\delta}$ (t), is integrated to identify the total area from ω = 0 to infinity, and the cutoff frequency at the current time instant is determined as the frequency where half the total area is captured.

$$\frac{\psi_1^2(t)}{\psi_{tot}^2(t)} = \frac{\frac{1}{2\pi}\int_0^{\omega_c(t)} G_{\delta\delta}(t)d\omega}{\frac{1}{2\pi}\int_0^{\infty} G_{\delta\delta}(t)d\omega} = 0.5$$

Equation 2: Calculation of time-varying cutoff frequency.

The cutoff frequency is scaled by the maximum power over the frequency range at the time instance under consideration. This is performed for all time instances to develop the power frequency (ω_G (t)), and is shown in Equation 3. Although Equation 3 is written for the time-varying case, it may be applied similarly using the discrete Fourier transform data to determine an average power frequency.

$$\omega_G(t) = \omega_c(t) \max_{\omega} G_{\delta\delta}(t)$$

Equation 3: Calculation of time-varying power frequency.

RESULTS

The results from the study are organized into two parts. First, overall correlation of the CHR, POWS, and TLX scales is examined. Second, observations from the data are discussed, with a focus on how the pilot's perceived the degraders they encountered in the testing.

Overall Correlation

A note on the data presented in the figures examining correlation that follow is warranted. Metrics, identified in Ref. 6 were computed based on the study data. Quality of correlation is desired between the CHR, POWS, and TLX scale to assess the ability of the quantitative metrics to serve as a measure of workload. To that end, some statistical analysis is performed. Most metrics are plotted along with the correlation coefficient computed assuming a best-fit line of the form $\hat{y} = \widehat{B_o} + \widehat{B_1}x$, and an associated prediction band. The coefficients of the best-fit line were selected to minimize the least square errors between the fit value and the data from an individual pilot. The 100(1-α) prediction band for the quantity $\widehat{B_o} + \widehat{B_1}x$ is given as:

$$\widehat{B_o} + \widehat{B_1}x \pm t_{n-2,\alpha/2} \cdot s_{pred} \text{ , where}$$

$$s_{pred} = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Equation 4: Prediction band calculations.

In Equation 4, $t_{n-2,\alpha/2}$ comes from the t-distribution as a function of the number of degrees of freedom in the dataset and the desired level of confidence in the prediction band, and s represents the standard deviation of the error. The prediction bands included in the following plots are for a 95% confidence level. Thus, for a given value of an evaluated metric, an estimate of the qualitative workload metric is provided by the corresonding value of the best-fit line, and a prediction interval is bounded by the prediction bands to a 95% confidence level.

The linear form of the best-fit line represents an assumption that the PIW metric increases linearly with workload. This assumption is potentially problematic, especially with the CHR scale considering the nonlinear, or perhaps piecewise linear, nature of the scale with regards to workload. The POWS and TLX scales may or may not be interpreted linearly when evaluations are given by the pilot. The use of a more complex fit line may turn out to be appropriate; however, there was a desire not to assume a complicated underlying model structure without first principles knowledge of the relationship present in the data. What was desired, was a simple interpretation to compare the correlation between the PIW metric and qualitative scale, and this is provide directly by the correlation coefficient.

Assuming a linear model form for qualitative handling quality ratings is not unprecedented. Ref. 12, for example, successfully demonstrated nearly monotonic increases in evaluated mean and max power frequency as a function of time with HQR rating for two pilots. Thus, the use of linear statistical methods provides rough measures of correlation to assess the ability of PIW metrics to stand in as an analog to qualitative workload evaluations.

The design of the study was influenced by the results and discussion presented in Ref. 6. The expectation going into the data analysis was that the POWS scale and perhaps the NASA TLX ratings would correlate to PIW measurements better than traditional CHR workload ratings due to how CHR ratings merge disparate assessments (workload and performance) in the process of generating an overall workload assessment. It is noted that NASA TLX also combines various aspects of workload, but for this study the relative weight of each input factor remains the same as the subject matter experts are exposed to degraders.

Overall correlation was first investigated using the most clinical task data available – the single axis sum-of-sines. The actual results have revealed that the correlation is highly dependent on many factors, as the results do not consistently indicate one scale is preferred compared to others.

Consider first single axis sum-of-sines results shown in Figure 9, generated by Pilot 1001. A legend to assist in interpreting the correlation figures is shown as Figure 10. Figure 9 plots the qualitative workload rating versus the PIW metric in various markers (representing specific nominal or degraded configurations), along with a bold best fit line and the dashed line representing the prediction band to a 95% confidence level. Reasonable, but perhaps not as strong as desired correlation is present between the PIW metric and all three qualitative workload ratings. The difference between HQR, POWS, and TLX results is not stark. In fact, approximately the same correlation coefficient and spread in the prediction band is present. Further, the quantitative data (aggressiveness in this case) is spread relatively evenly across a large band promoting confidence in the existence of the correlation.

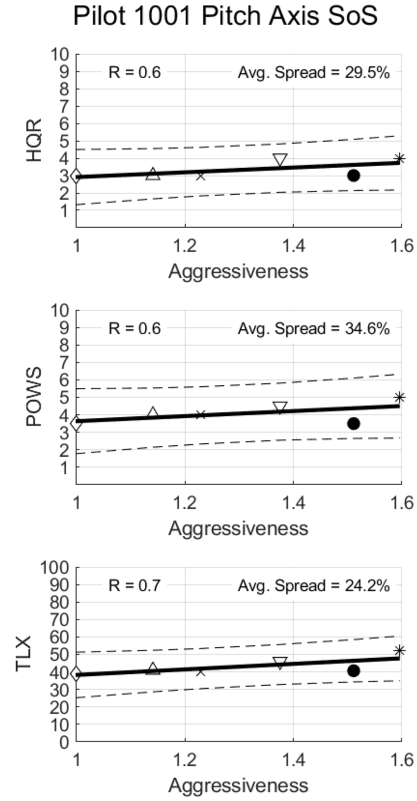

Figure 9: Pitch SoS Correlation with CHR (top), POWS (middle), and TLX (bottom).

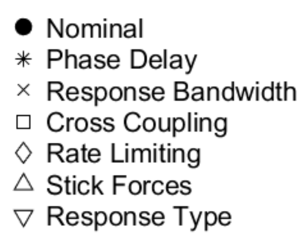

Figure 10: Legend for Interpreting Correlation Figures.

The similar character of correlation among qualitative workload scales is generally not apparent in other pilot data, conducting the same single axis sum-of-sines task. Consider the Pilot 1012 data, representing the most common interpretation of the data we collected in Figure 11:

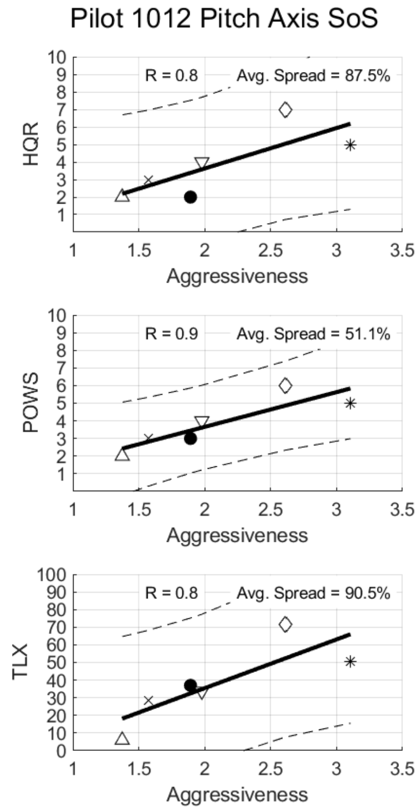

Figure 11: Pilot 1012, Pitch Sum-of-Sines Data.

The Pilot 1012 data shows the trends apparent in 8 out of 12 cases for pitch sum-of-sines task data – the POWS correlation with PIW metrics provides the lowest spread, therefore best predictive capability, followed by the CHR comparison. The predictive capability for the TLX data is worse than the other qualitative metrics for 11 out of the 12 pilot's data. It is not uncommon to see TLX predictive bands spreading over 75% of the TLX scale range, effectively nulling any predictive capability based on PIW metrics as the input. The poor predictive capability of TLX when associated with a handling qualities task combined with the fact that this scale is the most difficult to implement dim the prospect of using this scale in future experiments related specifically to handling qualities.

Overall, correlation measured by correlation coefficient was poorer than anticipated for the PIW metrics of aggressiveness and mean power frequency, with wide ranges seen across the pilot population. Table 4 summarizes the correlation coefficients computed by comparing the qualitative workload ratings against the aggressiveness metric for the pitch only SoS task. In general, most subjects aligned exhibited a positive correlation, albeit to varying degrees, which aligned with previous expectations. With closer inspection, several negative values appear which was unanticipated. One may be quick to dismiss the results of pilot 1010 since their flight experience was quite limited, however pilot 1011 exhibited similar traits and was an active Marine test pilot with over 1800 flight hours. This incongruence challenges the premise that all pilots materialize workload in a qualitatively similar fashion and demonstrates the challenge in seeking the universal application of a PIW metric.

Table 4: Correlation Coefficient Summary – Pitch SoS Task for Aggressiveness PIW Metric.

Evaluator	R - CHR	R - POWS	R- TLX
1001	0.6	0.6	0.7
1002	0.6	0.4	0.4
1003	0.6	0	0.1
1004	0.8	0.9	0.3
1005	0.8	0.8	0.7
1006	-0.3	0	0.4
1007	0.3	0.9	0.7
1008	N/A	0.5	0.1
1009	0.7	0.7	0.6
1010	-0.6	-0.6	-0.6
1011	N/A	-0.3	-0.5
1012	0.8	0.9	0.8

Following the results and conclusions from the first survey of PIW metrics (Ref. 6), it was identified that correlation would likely need to be specific per-pilot, or at least to "types" of pilots. The hypothesis was that with data collected for purpose coupled with fitting each, individual pilot, a sufficiently strong correlation could be established to allow application of PIW metrics as a more regular tool during HQ evaluations in lieu of subjective ratings. The data suggests that, while it seems that considering individual pilot correlations is necessary, it is not sufficient with PIW metrics. The results show that some individual pilots demonstrated strong correlation, while some pilots demonstrated no identifiable correlation at all. This does not necessarily indicate that a given pilot is objective in their assigned workload ratings. It only means that the workload perceived by some pilots don't significantly overlap with the quantitative measures of workload attempted to be captured by the PIW metrics. It suggests that different metrics should be investigated as a standalone, or combinatorial measure of workload perhaps relying on additional data, such as physiological measures, which were not captured in this study.

Effect of the Degraders – Sum-of-Sines Tasks

In a perfectly designed experiment, the baseline CHR configuration would have always received a Level 1 qualitative rating, meaning the pilot was able to achieve desired task performance by modulating perceived workload. Pilot 1012's ratings are selected to highlight how close to this ideal the experiment came.

The baseline model characteristics were designed using the quantitative criteria in ADS-33E to provide borderline level 1/level 2 HQ. For the single-axis SoS MTEs, this was found largely to be the case as shown in Figure 12. When examining this figure, and similar examples that follow, the qualitative ratings for each degrader are shown as black bars, with the nominal configuration indicated by the red line. The intent is that this allows a quick look at relative difficulty, as perceived by the pilot. That said, there was a marked increase in both HQR and POWS associated with the multi-axis SoS MTE as show in Figure 13. Not only does the increased task complexity increase the target baseline, but it also seems to exaggerate the effect of the degraders. Potentially because the pilot is starting to run out of capacity to deal with degraded modes and the harder task (i.e. the pilot is becoming task saturated).

For the majority of the pilots, the degraders tended to result in worse HQRs and higher workloads than the baseline configuration. For a couple of the pilots, this was not the case as exemplified in Figure 14. This could be due to a learning effect as the pilots were shown the baseline configuration first (though they did not know it was baseline). This learning effect occurred even though pilots were encouraged to take as many practices runs as necessary to feel competent in the task.

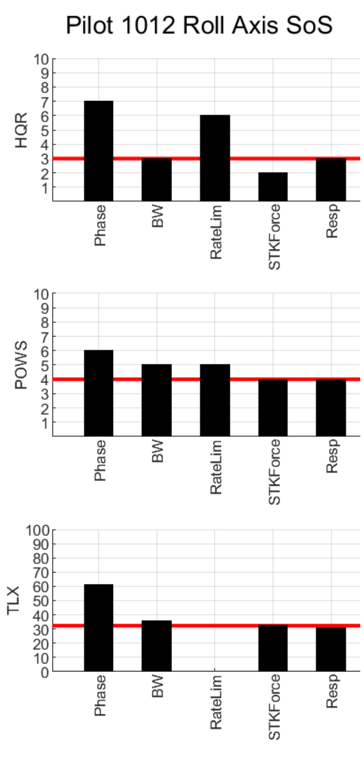

Figure 12: Pilot 1012 Roll Sum-of-Sines Pilot Ratings.

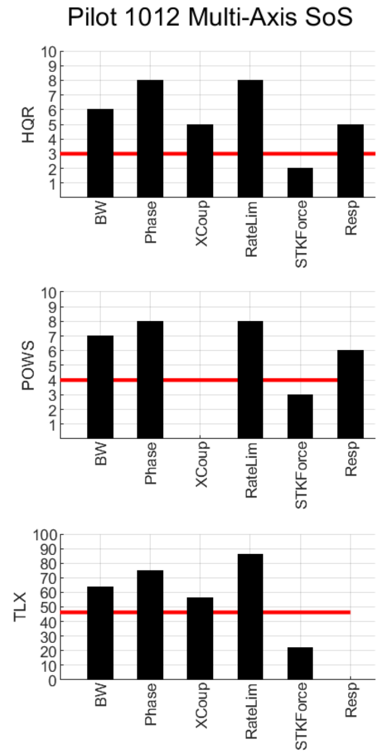

Figure 13: Pilot 1012 Multi-Axis Sum-of-Sines Pilot Ratings.

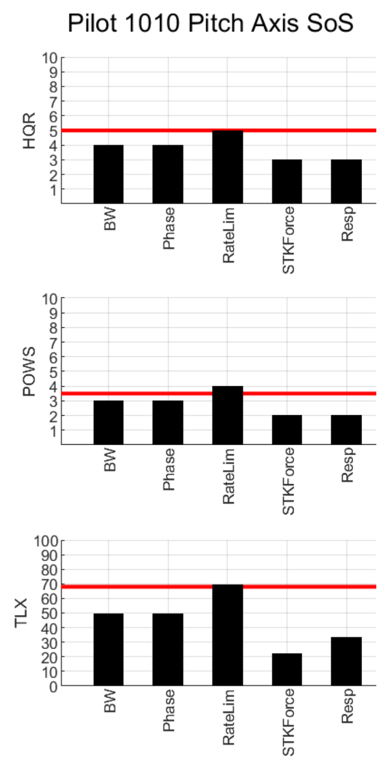

Figure 14: Pilot 1010 Qualitative Ratings for Pitch Axis SoS.

For the case of the attitude command system, the effect of the model change was intended to improve the handling qualities by augmenting the traditional rate command system. Interestingly, a majority of pilots found this configuration more objectionable, as shown in Figure 13. The pilots often made comments that the sensitivity was too high which caused them to over control the aircraft. One potential cause for these responses may include the pilot's adaptation to a different response type, in which case the study accommodated for this by permitting pilot's to conduct proficiency run prior to scoring, although subjects rarely capitalized on this opportunity. A second cause may have been the gearing ratio between the inceptor and attitude output.

An expected result is noted when considering the relative location of the nominal configuration and phase delayed configuration in terms of the PIW metrics and pilot ratings. In a closed-loop compensatory tracking task, the phase delay is expected to result in higher PIW measurements due to the need to over apply the input to illicit an initial response and then continue to over control to drive the response to the desired set point. Most often when considered on a per-pilot basis, the qualitative and quantitative data did show this expected trend. A sample case is seen in Figure 15, where the phase delayed configuration is in the upper right-hand corner of the plot for all qualitative scales, maximizing the PIW metric of mean time-varying power frequency in the longitudinal control axis.

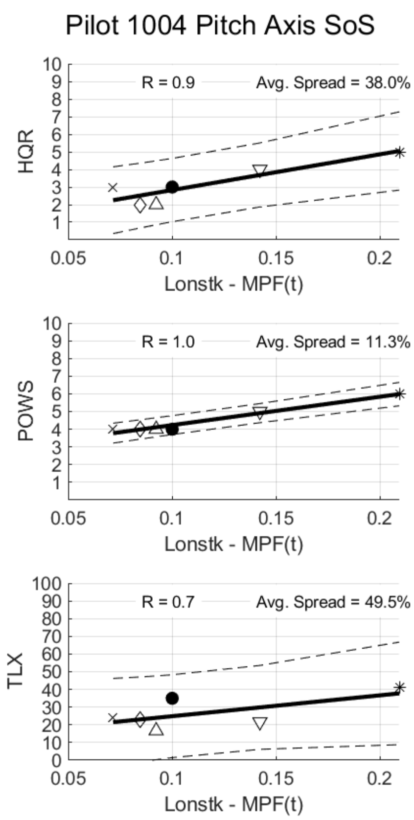

Figure 15: Pilot 1004 Exhibiting Max Qualitative/Quantitative Workload in Phase Delayed Configuration.

Effect of the Degraders – Precision Hover Task

Elaborating on the expectation that the baseline model was Level 1, it stands to reason that the first five degraders (Stability, Phase delay, Bandwidth, cross couplings, and rate limit) would yield ratings that are higher than baseline. Furthermore, the modified response type was expected to reduce pilot workload thus leading to ratings equal to or lower than baseline. Increasing the stick force was initially expected to degrade handling qualities, but in the course of collecting data, it was generally considered by pilots as an improvement as it provided a sense of damping when they might otherwise tend to over control the aircraft. The precision hover was the most challenging of the five tasks, so much so that two pilots were incapable of completing it for record. Of the remaining pilots that were successful, two pilots rated the baseline at Level 1, six pilots as Level 2, and finally two pilots as Level 3. Figure 16 shows the precision hover results for Pilot 1001 which best exemplifies the expected shape of the data.

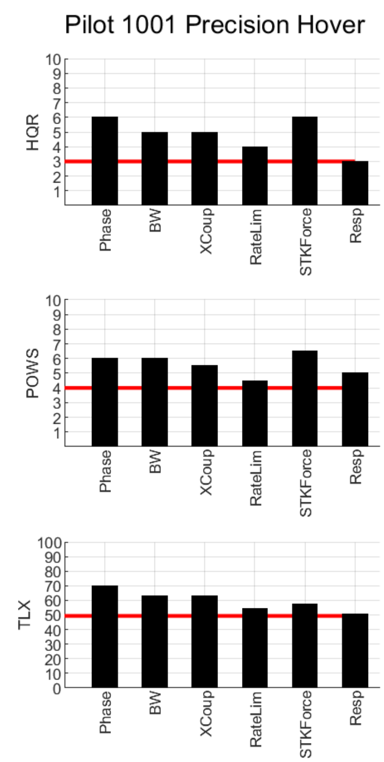

Figure 16: Pilot 1 Data shape Exemplified.

The modified response type was expected to reduce pilot workload, and it was generally an improvement resulting in an HQR reduction of approximately 2 from the baseline assessment. This was in contrast to the SoS result in which the modified response type did not markedly reduce workload. This inconsistency can be attributed to the fundamental difference in the style of tracking employed for the Precision Hover MTE as compared to the SoS task.

A well know challenge in the pursuit of assessing handing qualities is achieving a true assessment at the hands of a focused, proficient, and competent pilot. The learning curve associated with undertaking an unfamiliar task must be managed so as prevent unfairly inflated pilot ratings. Despite best efforts, some pilots insisted on their readiness to collect data before reaching full proficiency. This is demonstrated in the results for pilot 11 which are shown in Figure 17. Their baseline assessment was exceptionally high on all workload scales. Subsequent events in which the degraders were applied indicate a slight reduction in the HQRs but overall still high workload according to the alternate scales. Performance peaked during the assessment of stick force and response type, initially breaking into the level 2 category before ultimately giving an HQR 4 rating and achieving level 1 standards. This marked improvement was the synthesis of repetition and favorable model settings.

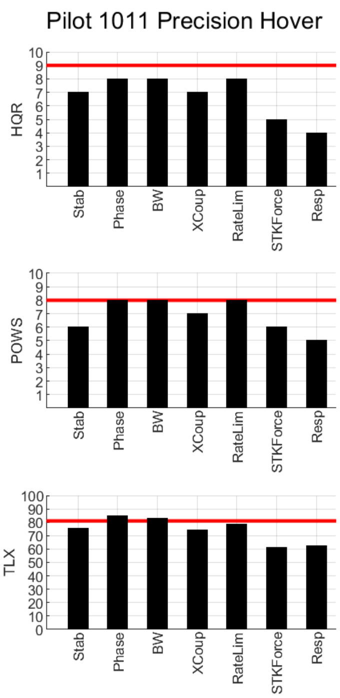

Figure 17: Pilot 11 gaining proficiency through repetition.

Overall, the workload ratings across all scales investigated were generally consistent for the population tested. Pilot 7 however presented with somewhat conflicting ratings, in particular between HQR and POWS, as seen in Figure 18. Their HQRs averaged approximately 4 while their POWS ratings were near 7. This pilot found themselves in a quandary as they were routinely working extremely hard and achieving desired performance. In practice, it's far more typical to observe pilots complacent to achieve adequate tolerance at a much lower workload and thus motivate them to seek desired performance. In this pilot's situation, it was the structure of the HQR scale that drove them to select a rating consistent with their performance level even though the rating failed to capture the workload necessary to reach that threshold. This fundamental constraint of the HQR scale was the motivation to explore the suitability of alternate workload scales that may better correlate PIW metrics to perceptions of workload.

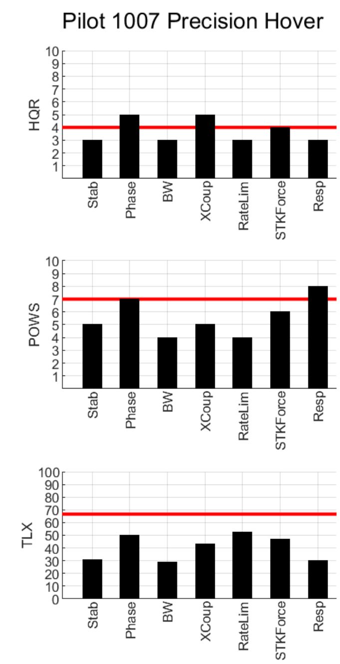

Figure 18: Pilot 7 Rating variation across scales.

Effect of the Degraders – Reposition Task

For the reposition task, only POWS and TLX qualitative rating data was collected due to the lack of defined tolerances for the open-to-interpretation task. The reposition task data did not add value to the attempt to find correlation between the PIW metric and qualitative handling quality rating.

Of note, we saw little variation in the nominal configuration workload for the evaluation pilots, where POWS range remained quite limited from 2-4. The effect of the degraders was more pronounced for some pilots than others. Some pilots kept POWS ratings from 2-6 for all degraders, whereas in the extreme case we saw Pilot 1009 rating the bandwidth degrader a POWS 10, although this did not comport with the corresponding TLX rating of 63, which was exceeded by other degraders on the TLX scale.

Experimental Data Collection Lessons Learned

In executing the experiment, several observations were noted that would greatly benefit future researchers interested in recreating this type of experiment. The criticality of the simulator field of view is a function of the task to be completed. Two of the tasks in this experiment required the pilot to fly typical dynamic helicopter maneuvers. While the reposition task had only a few key requirements to meet, the hover MTE was highly specified in time and distances and required extreme precision to perform to the desired tolerances. The limited field of view of the single seat simulator at USNTPS, which was an approximately 8 feet by 8 feet flat projection on a wall approximately 10 feet in front of the pilot, did not permit the subject to view any of boresight cues critical to accomplishing the Hover MTE which severely restricted their ability to accurately perform the task and assess quality. Guidance from the test conductor on how to use the available cueing off the nose to assist in managing fore/aft position was used, however this cue was highly sensitive to the air vehicle pitch attitude.

Recurring pilot comments such as "I can't make an input small enough" and "I feel a heavy clunk as I move the cyclic stick through trim" highlighted an influence of the flight control system on the pilot's ability to perform some tasks. An audible squeal in the roll axis also indicated that potential bearing wear and friction in the control mechanism was contributing to the overall experience. It was apparent that the flight control system mechanical characteristics were having an influence on the pilot workload with the effect most apparent during high precision tracking when stick amplitudes were very small as in the positon keeping portion of the precision hover MTE or certain portions of the SoS task. Given that all subjects were exposed to the same flight control system, the data are expected to be comparable across the sample group, although this variable was not an intended source of degradation to inflate pilot workload. Correlation is expected to still be viable, although the magnitude relationship of the metric to the qualitative rating may be biased higher than if the flight control system was optimally designed.

Pilot Proficiency

Observations from the simulation sessions overwhelmingly substantiate the notion of a proficiency curve with respect to conducting a task. While the pilot was afforded the opportunity to perform each new task to become familiar, it was typical for them to accept only one or two practice runs prior to taking data for score. This element did not factor heavily in the simplified SoS tracking, but became strongly apparent with the much more challenging reposition and hover tasks. The learning curve can be discretized into three sections. In the first section, the pilot is translating their understanding of the task into control inputs to perform the task. Task performance typically improves demonstrably after 3-5 attempts at which point it plateaus. This period continues until fatigue and/or frustration causes performance to degrade with further attempts. Each pilot's curve is unique, but the overall trend appears consistent. The pilot's reluctance to conduct additional proficiency runs prior to taking workload data for the complicated tasks meant the pilot was still climbing the initial learning curve before reaching their individual plateau and resulted in artificially inflated CHR. While this alone may not be significant, the order in which the degraders were presented to the pilot will need to be carefully examined to prevent false conclusions that the degraders presented early in the sequence were in fact harder to fly when the reality may be that pilots were still gaining proficiency through those runs. This was verified in one pilot's case who, when complete with all degraders for the hover MTE, was asked to repeat a previously objectionable case in which vehicle control was in question and performed the task safely. The scope of data collection averaged 3 to 4.5 hours per subject and likely contributed to some pilot's unwillingness to accommodate their natural learning curve for each task.

CONCLUSIONS

For the purposes of correlation with PIW metrics, it seems that neither CHR nor POWS was notably better than one another. This was surprising since CHR is not solely a workload scale whereas POWS was intended to be.

For the purposes of correlation with PIW metrics, it seems that TLX is consistently less correlatory than either CHR or POWS. This does not necessarily mean TLX isn't an appropriate tool for evaluating workload in an HQ sense, just that it doesn't emphasize the form of workload being captured by PIW metrics.

It is becoming more evident that establishing quantitative correlation of workload will require consideration of pilots as individuals but that, at least for PIW metrics, not all pilots can be expected to show strong, or even any, correlation.

In cases where the most offending degraders did not illicit objectionable workload, pilot control strategies did not sufficiently stress the system which reinforces the importance of using multiple pilots to assess the suitability of a flight control system.

Future experimental design should consider the end to end performance of the system and the desired effect verified by a test pilot.

Future research should consider reducing scope, breaking sessions into multiple events, and forcing additional practice runs to fully develop pilot proficiency prior to collecting data.

Author contact: Ryan Paul ryan.paul@okstate.edu
Matthew Rhinehart matthew.h.rhinehart2.civ@us.navy.mil

Kenneth Cahill kenneth.m.cahill2.civ@us.navy.mil
Zachary Smith Zachary.f.smith6.civ@us.navy.mil

ACKNOWLEDGMENTS

R. C. Paul thanks the ONR Summer Faculty Program for the support provided to perform the analysis portion of this work. The team would like to thank the NAWCAD NISE Office for being the primary funders and backers of this research. In addition, a thank you to all those involved in the simulation test; especially the pilots who performed the flight tests and US Navy Test Pilot School for the use of their facilities.

REFERENCES

[1] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,," *Advances in Psychology,,* vol. 52, pp. 139-183, 1988.

[2] R. Paul and M. Rhinehart, "Exploring Pilot Workload Using Inceptor Time Histories," in *Vertical Flight Society 76th Annual Forum and Technology Display*, Virtual, 2020.

[3] A. Lampton and D. Klyde, "Power Frequency: A Metric for Analyzing Pilot-in-the-Loop Flying Tasks," *Journal of Guidance, Control, and Dynamics,* vol. 35, no. September-October 2012, pp. 1526-1537, 2012.

[4] G. E. Cooper and J. R. P. Harper, "The Use of Pilot Ratings in the Evaluation of Aircraft Handling Qualities," National Aeronautics and Space Administration, Washington, D.C., 1969.

[5] D. Wilson and D. Riley, "Cooper-Harper Pilot Rating Variability," AIAA Atmospheric Flight Mechanics Conference, Boston, MA, 1989.

[6] D. Riley and D. Wilson, "More on Cooper-Harper pilot rating variability," in *17th Atmospheric Flight Mechanics Conference* , Portland, OR, USA, 1990.

[7] United States Army Aviation and Missle Command, *ADS-33E-PRF Aeronautical Design Standard Performance Specification Handling Qualities Requirements for Military Rotorcraft,* Redstone Arsenal, AL: Army Aviation Engineering Directorate, 2000.

[8] C. Hanson, J. Schaefer, J. J. Burken, D. Larson and M. Johnson, "Complexity and Pilot Workload Metrics for the Evaluation of Adaptive Flight Controls on a Full Scale Piloted Aircraft," National Aeronautics and Space Administration, Dryden Flight Research Center, Edwards, CA, 2014.

[9] E. J. Field and S. E. Giese, "Appraisal of Several Pilot Control Activity Measures," AIAA Atmospheric Flight Mechanics Conference, San Francisco, CA, 2005.

[10] S. Jennings, G. Craig, S. Carignan, K. Ellis and D. Thorndycraft, "Evaluating Control Activity as a Measure of Workload in Flight Test," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* vol. 49, no. 1, p. 64–67, 2005.

[11] M. B. Tischler, Aircraft and Rotorcraft System Identification, Reston, VA: American Institute of Aeronautics and Astronautics, Inc., 2012.

[12] S. A. Conner and W. W. Wierwille, "Comparative Evaluation of Twenty Pilot Workload Assessment Measures Using a Psychomotor Task In A Moving Base Aircraft Simulator," National Aeronautics and Space Administration, Moffett Field, CA, 1983.

[13] I. Niewind, "Pilot Gain and the Workload Buildup Flight Test Technique: A Closer Investigation of Pilot Inceptor Workload," German Aerospace Center, Braunschweig/Manching, 2012.

[14] I. Niewind, "Pilot Gain and the Workload Buildup Flight Test Technique: A Closer

Investigation of Pilot Inceptor Workload," German Aerospace Center, 2012.

[15] D. Klyde, S. Pitoniak, P. C. Schulze, P. Ruckel, J. Rigsby, C. Fegely, H. Xin, W. Fell, R. Brewer, F. Conway, R. Mulato, J. Horn, C. Ott and C. Blanken, "Piloted Simulation Evaluation of Tracking Mission Task Elements for the Assessment of High-Speed Handling Qualities," *Journal of the American Helicopter Society,* vol. 65, no. 032010, pp. 1-24, 2020.

[16] D. T. McRuer, D. Graham and a. I. Ashkenas, Aircraft dynamics and automatic control, Princeton University Press, 2014.

[17] W. Navidi, Statistics for Engineers and Scientists, New York, NY: McGraw Hill, 2008.