OPERATOR STATE MONITORING FOR WORKLOAD PREDICTION AND MANAGEMENT

Dr. Martine Godfroy-Cooper, martine.godfroy@sjsu.edu, SJSU/ NASA ARC, Moffett Field, CA, USA; Joel D. Miller, SJSU/ NASA ARC, Moffett Field, CA, USA; Edward. N. Bachelder, SJSU/ NASA ARC, Moffett Field, CA, USA; Zoltan Szoboslay, U.S. Army DEVCOM AvMC, NASA ARC, Moffett Field, CA, USA

Abstract

Ongoing efforts to modernize the U.S. Army is resulting in the development of the next generation of rotary aircraft under the Future Vertical Lift (FVL) program. FVL missions will be characterized by increased agility, degraded visual environments (DVE) and optionally piloted vehicles (OPVs) in complex, highly contested and dynamically changing environments. New technologies and automation in the cockpit allow aviators to be more effective in completing their missions in the modern dynamic battlefield. Autonomous systems have the potential to contribute to improvements in operational safety, efficiency and effectiveness but have also introduced additional human factors engineering considerations. These technologies provide pilots with appropriate and timely information and support, while avoiding overloading with excessive clutter and information. At the same time, excessive automation can lead to overloading/ underloading, leading to automation misuse, complacency, and loss of situational awareness (SA).

Adaptive decision aiding systems have the potential to enhance operators' capabilities by recognizing situations, external (environment) and internal (operator state) and providing subsequently real-time adaptation of the aircraft controls and human/ machine interfaces. A pre-requisite to the development of such adaptive decision aiding systems is the real-time monitoring and identification of the operator's performance and mental state in terms of workload and situation awareness. Being multidimensional, latent factors, workload (WL) and SA need to be inferred through observable variables such as subjective assessments, task-based metrics, and psychophysiological measures. Of particular interest for FVL are techniques and technologies that enable real-time WL and SA assessment, for realistic missions in simulator and/or real flight conditions.

The present research aims to provide a methodology to manipulate WL and SA in an operationally relevant mission in a fixed-platform UH60 simulator, evaluate the usability, diagnosticity, sensitivity and reliability of various performance metrics in situ using novel approaches for the data analysis and validate real-time predictive models of WL. An experiment was designed to manipulate WL and SA by using two levels of flight control level of automation (LOA) and two levels of obstacle cueing symbology during a degraded visual environment (DVE) medical evacuation (MEDEVAC) mission. The level of WL was further manipulated by the presence of obstacles and aircraft survivability equipment (ASE) threat events (radar and missile warnings). Five UH60M Army experimental test pilots participated in the simulation at the U.S. Army rotorcraft in flight laboratory RIFL system integration laboratory (SIL) at Fort Eustis, VA.

Preliminary analyses validate the WL driver's selection, and the usability of the real-time subjective WL report using a modified Bedford rating scale, used to compute the spare capacity operations estimator (SCOPE). A novel approach to statistical analyses is proposed, that will allow to determine the relative weight of each metric to the final WL and SA estimates, and support performance modeling.

1. INTRODUCTION

Ongoing efforts to modernize the U.S. Army is resulting in the development of the next generation of rotary aircraft under a program known as Future Vertical Lift (FVL). FVL missions will be characterized by increased agility, degraded visual environments (DVE) and optionally piloted vehicles (OPVs) in complex, highly contested and dynamically changing environments.

New technologies in the cockpit of FVL aircraft and upgraded current fleet aircraft allow U.S. Army aviators to be more effective in completing their missions in the modern dynamic battlefield. These technologies include advanced ownship sensors (IR, LIDAR, RADAR), synthetic vision systems (SVS) for operations in DVE, video feeds from other team players (UAVs, team piloted aircraft, team ground vehicles), interconnected tactical maps, encrypted text communications, automatic flight control systems

(AFCS), and advanced aircraft survivability equipment (ASE).

Autonomous systems have the potential to contribute to improvements in operational safety, efficiency and effectiveness but have also introduced additional Human Factors Engineering (HFE) considerations to the design of the associated Human-Machine Interfaces (HMIs) and Interactions. These technologies provide the pilot with appropriate and timely information and support, while avoiding overloading with excessive clutter and information. At the same time, excessive automation can lead to mental and/or physical overloading or underloading, leading to automation misuse, complacency, and loss of situational awareness (SA).

Current research concepts envision the use of adaptive aiding systems which have the potential to enhance operators' capabilities by appropriate adaptation of the HMI. These adaptive aiding systems are capable to recognize situations, external (environment) and internal (operator state) when the operator requires assistance and provide a real-time adaptation to the new constraints.

The objective of the Holistic Situational Awareness-Decision Making (HSA-DM) Program is to develop technologies that enable sustainable levels of task loading for FVL pilots (both the pilot-on-the-controls, and the air mission operator) by developing and transitioning cognitive workload management capabilities. Intelligent pilot information management systems will consider the current situation, the current workload level of the pilots, expected increases/decreases in workload based on mission phase and operator cognitive state (workload, SA, engagement). The pilot's cognitive states in situ will be used to drive adaptation in HMIs, which will provide optimized task loading for FVL warfighters to improve combat mission performance in all conditions.

The technologies identified to optimize task loading include:

- Advanced Flight Controls: Improved stability, with automatic and autonomous modes. Tactile feedback through stick soft stops and force feedback.
- Human Machine Interfaces: HMIs technologies that facilitate intuitive communication of knowledge to the pilots relevant to the current situation and mission. Make use of multi-modal pilot cueing including visual, spatial-auditory, and tactile cueing. Improve pilot input devices including touchscreens and voice recognition.
- Autonomous Decision Aiding (DA): Hard and soft fusion from all sources of information on-board and from remote sources, situational awareness world model database, context sensing and awareness, and decision aiding algorithms.

A pre-requisite to the development of such adaptive decision aiding systems is the real-time monitoring and identification of the operator's performance and mental state in terms of workload and situation awareness. Being multidimensional, latent factors, workload (WL) and SA need to be inferred through observable variables such as subjective assessments, task-based metrics, and psychophysiological measures (biometrics). Of particular interest for FVL are techniques and technologies that enable real-time WL and SA assessment, for realistic missions in simulator and/or real flight conditions.

Attempts have previously been made to measure either workload or SA, often with instruments or methods that were not enough sensitive to the dynamics of human performance during the task, leading to confusion about which construct was really measured (WL or SA), and how the factors underlying these constructs interact. Other limitations include the lack of ecological validity of the simulation, the task and population idiosyncracies. However, the drawback to high-fidelity simulation is that the collected quantitative data are usually very noisy, due to the length of the experimental trial (sometimes up to 30 minutes) and the lack of control over the environment induced by pilots performance variations, such as ownship heading, speed and altitude. Therefore, analysis of performance metrics over an entire run will be contaminated by uncontrolled variables and the total explained variance due to the controled varaibles, such as the induced levels of WL and SA, is usually to low to lead to significant results.

The objectives of the present study were: 1) to select and collect relevant measures of WL and SA for the development of real-time OSM; 2) to develop a methodology for the analysis of time series data, in order to identify the local WL triggers, quantify the contribution of each metric to the assessment of WL and SA, evaluate each factor communality between measures within and between measures; and 3) propose foundational elements for the modelling of the relationships between WL, SA, and operative performance.

1.1. Articulating Workload, Situation Awareness, and Performance

Human performance is influenced by internal and external factors. Internal factors include motivation, experience and skills, perceived workload, attention and vigilance, situation awareness and teamwork ability [Ref [1]]. External factors include task definition, training, task distribution, as well as HMI, automation and support tools). Until recently, most of the research has focused on the impact of a single factor (such as workload or fatigue) on human performance. However, mishaps often result from the interaction of multiple factors and this interaction is still mostly

underexplored. There is a need to determine the key factors and their associated behavioral markers so that either the pilots themselves can recognize that their performance is under threat and compensate accordingly, or there should be automated support to augment human performance or take over control.
Two aspects of human performance, WL and SA (also enabling indicators of attention and vigilance) have been selected for the present study as they appeared being the factors with the highest impact on pilots' performance as well as the most relevant measures to consider in the framework of the HSA-DM program. Although other factors such as fatigue, stress or trust in automation were also considered as important, the ability to obtain reliable indicators of these states remains challenging.

1.1.1. Cognitive Workload

To date, and despite a growing interest in the topic, there is still no universally accepted definition of the the concept of workload, and that the term itself did not appear until the 1970's [Ref.[2]]. The term was first introduced in the framework of theories of attention and performance, and identified as a critical factor in human performance and more generally, in system effectiveness. Over the years, the definitions have evolved, and workload has been associated with imposed task demands, the level of performance an operator can accomplish, the mental and physical efforts an opeartor exerts, and the operator's perception. If high WL is associated in general to performance decrement, the opposite is also true, and low cognitive WL can also have detrimental effects on performance [Ref. [3]].
The operational definitions of workload from various fields continue to show discrepancies regarding workload sources, mechanisms, consequences, and measurement. In general aspects of workload have been categorized between: 1) the amount of work and number of things to do; 2) the time and the particular aspect of time one is concerned with; and, 3) the subjective psychological experiences of the human operator [Ref. [4]]. One of the reasons why the concept remains elusive is that workload is a global multidimensional, multifaceted construct, resulting from the aggregation of many different demands and reflecting the interaction of these demands imposed on operators by tasks they attend to [Ref. [5]]. Because it is not directly observable, it must be inferred from observation of overt behaviour or measurement of psychological and physiological processes. For these reasons, it has also been referred to as a latent factor or an intervening variable by Gopher and Donchin [Ref. [6]]. The authors thought that no single, representative measure of workload exists or is likely to be of general use, although they do not provide guidance on how many workload measures they feel are necessary or sufficient.

1.1.2. Situation Awareness

Situational awareness (SA) is also a global construct, that can be defined as "knowing what is going on around you" [Ref. [7]] and more formally, as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future". SA is theorized to be critical to human performance, and appears to account for as much as 88% of human error [Ref. [8]]. a more recent and contextual definition is provided by Chauvin [Ref. [9]] that describes sa as mental model that when optimal, drives appropriate diagnostic and decision-making. meanwhile, degraded sa can lead to a logical decision for the operator, but inappropriate to the real situation and can lead to accident.

1.1.3. Interactions between Workload, Situation Awareness, and Performance

Because optimal mental workload, situation awareness (SA), and effective decision making are mission critical [Ref. [10]], it is important to model the many facets of the constructs and how they inter-relate. For example, Hendy [Ref.[11]] proposed a model of human operator information processing in which both workload and SA are direct ly influenced by the availability of processing resources and time availability, hypothesis partially driven by Wagner's Limited Capacity Theory (LCT), [Ref. [12]].
Mental WL and SA are considered inter-related yet independent constructs [Ref. [13]]. Endsley pointed out that plausible cases could be made for any combination of extremely high or low SA and extremely high or low WL. Consequently, it is impossible to predict the level of one from the other. The lack of connection appeared to become even more convincing when Endsley considered the relationships of both mental WL and SA with performance. In both cases, the concepts have been hypothesized and demonstrated to dissociate from the quality of performance, under some conditions.
For other authors, the relationship between mental WL, SA and performance can take the form of a causal and logical relationship connecting WL with SA and SA with performance. For example, an increase in mental workload (a more demanding task) could eventually lead to a decrease in SA, which, in turn, could lead to inferior performances [Ref. [14], Ref. [15]]. The interactions between WL, SA and performance will vary as taskload, operator state, mission task elements, and environmental factors change over time.

1.1.4. Measuring Workload, and Situation Awareness

The multidimensional nature of mental WL and SA is reflected in the variety of metrics available. In general, measures of WL can be categorized as performance-based, or linked to the process of subjective self-

assessment, or associated with psychophysiology or neurophysiology. Each category has specific strengths and weaknesses [Ref. [16]] and the sensitivity of each measurement type can vary depending on the level of WL experienced by the operator [Ref. [17]].

Observed and Unobserved Variables

When confronted to variables that are not directly observable (unobserved variables), like WL and SA, it is necessary to identify observable variables that can provide information about these so called latent factors. In the example provided in Figure 1, WL is measured with five observable variables, and SA is measured with three observed variables. One can also see that some of the observed variables (electroencephalogram, EEG and eye-tracking data) are hypothesized to reflect both WL and SA, while conversely, NASA-Task Load indeX is expected to only reflect WL.

Assesment techniques

Three broad techniques have been traditionally used for assessing human performance and provide observable variables: subjective assessments, task-based metrics, analytical models, and psychophysiological measures.

1. Subjective assessments require the subject to assess his/her own performance through questionnaire-based feedback such as rating scales or structured interviews.
2. Task-based metrics measure the subject's performance using parameters related to either primary or secondary task performance. The primary task refers to the main task, while the secondary task refers to an additional loading task performed on top of the primary task. Performance-based methods are indirect measures of cognitive workload. They assume that an increase in task difficulty results in deterioration of performance, which increases the pilot's cognitive workload (or reduces the working memory capacity).
3. Psycho-physiological measures (biometrics) typically employ sensors for collecting the subject's physiological data, which are then translated into measures of performance using established psycho-physiological and analytical models. Among them, neurophysiological activity recording and interpretation is used for the assessment of the operator's mental state. Mental state monitoring can be particularly useful where more traditional methods have clear disadvantages.

2. METHOD

2.1. The selected measures of WL and SA

The following measures have been selected for their relevance in the context of airborne military operations or simulator studies:

- Subjective measures of WL [online five-point rating scale (real-time probe), NASA task load index (TLX) (post-trial self-rating), crew status survey (CSS)] and SA [SART (post-trial self-rating)], post-run and post-experiment questionnaires and free-report.
- Task-based metrics [reaction time to events, obstacle clearance (lateral, vertical), technical performance measures (lateral, vertical and heading deviation from flight path, speed deviation, radial error for Hover), pilot controls inputs (Spare Capacity Operator Estimate, SCOPE, Dynamic Interface Modeling and Simulation System-Product Metric, DIMSS-PM, Control Attack metric).
- Psycho-physiological measures (biometrics): Electro-Encephalogram (EEG) for the central nervous system (CNS), electro-cardiogram (ECG) for the autonomous nervous system (ANS) and eye-movements/ pupillometry for the peripheral nervous system (PNS). The selected variables are detailed in Table 1.

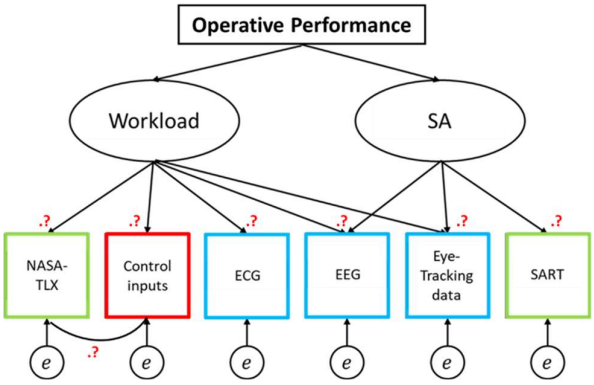


Figure 1: Example of a confirmatory factor analysis (CFA). Workload (WL) and situation awareness (SA) are the latent factors, depicted graphically with ovals, observed variables are graphically depicted by squares, with a different color for each variable category, *e* depicts measurement errors. Green: Subjective measures, Red: performance-based measures, Blue: Psycho-physiological measures.

Table 1. Selected Measures of WL and SA.

Subjective measures of Workload1, SA2	Bedford rating scale1 (post-trial self-rating), online modified Bedford rating scale1 (Real-time probe), NASA Task Load Index (TLX)1 (post-trial self-rating), SART2 (post-trial self-rating), Crew Status Survey (CSS)1 (post-trial self-rating)
Task-based measures (Domain-Specific)	Reaction time to events (obstacles and ASE events, WL prompts), technical performance measures (lateral, vertical and heading deviation from flight path, speed deviation, ability to recover intended position/speed control, radial error for Hover, SCOPE, DIMSS-PM, CPF, DC and Aggressiveness, CCM)
Analytical techniques	Pilot Control Models (a posteriori)
Psycho-physiological measures	Electro-Cardiogram [ECG: Heart Rate (HR), Heart Rate Variability (HRV)], Electro-encephalogram (EEG), Eye-tracking, Pupillometry

Operationalization of WL and SA in a relevant Mission Environment

2.1.1. Defining the WL and SA drivers in a MEDEVAC Mission

There are two ways to manipulate the human performance envelope (HPE) factors depending on the nature of the WL and SA factors, global or local.

Global factors are factors that remain constant throughout an experimental condition, such as the information display modality (Visual, Auditory, Tactile, Multimodal), the symbology sets (flight symbology, cueing symbology), the sensor feeds (IR, LIDAR, RADAR), the aircraft dynamic, the level of automation and the operational environment (GVE, DVE).

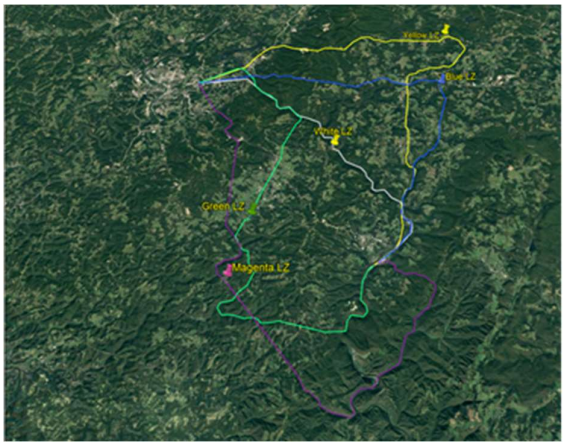


Figure 2. The five routes, one training (White route), four experimental (Blue, Magenta, Yellow and Green routes).

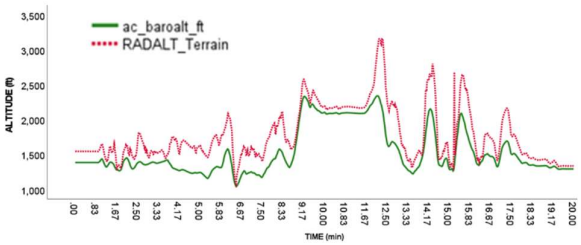


Figure 3. Terrain profile (green line) and flight profile (dashed red line) for the White route (training route).

Table 2. Global and Local Factors (red = used).

Global Factors	Local Factors
• DVE (constant) • IR (constant) • Obstacle and Hostile Threat Cueing: OFF, ON • Level of Automation (LOA): Mission Adaptive Autonomy Uncoupled, Mission Adaptive Autonomy Coupled	• Obstacle and Hostile Threat Events (12 per route), randomized between routes and conditions • Mission Task Elements (MTEs): Phases of Flight: Take-off from hover, Enroute, Approach to hover, Hover (order constant)

Local factors are transient events that occur several times during an experimental route, such as the presence of obstacles, hostile fire and/or precisely defined maneuvers referred to as Mission Task Element [MTE, Ref. [18]].

Five routes were designed (one for training, four for data collection) using a West Virginia database. The different route profiles were chosen to show identical levels of difficulty and duration. All routes initiated at

the Morgantown Airport and terminated at the Camp Dawson Airfield and largely followed terrain features such as valleys and roads (see Figure 2). The flight time varied between 23 to 30 minutes at the target speed of 100 kts and an average speed of 80 kts. The terrain profile for the training route (White Route) is depicted in Figure 3.
The Global and Local factors used in the present study are summarized in Table 2. Here, the visibility level and the sensor feed remained constant (DVE) throughout conditions.

a. The Global Factors

In the present experiment, two global factors were manipulated, obstacle cueing and level of autonomy (LOA), each factor having two levels:

1) Obstacle Cueing: The Integrated Cueing Environment Collision and Avoidance System (ICE-CAS): ON, OFF
2) Flight Control (FC) LOA: via the Mission Adaptive Autonomy (MAA) system: ON, OFF

Obstacle Cueing: ICE-CAS

The present simulation is capitalizing on the development of the ICE-CAS trimodal (visual-auditory-tactile) obstacle cueing suite developed in the context of the Degraded Visual Environment-Mitigation program (DVE-M) [Ref. [19]]. The ICE-CAS obstacle cueing suite allows 360° situational awareness about the ownship using a simulated bumper radar scanning 360° in azimuth in 4° increments at elevations -12°, 0°, and +12 (see Figure 4). The obstacles, man-made and terrain, detected by the radar are ranked in "urgency" (risk level) to determine the two most-urgent hits for cueing, giving rise to a risk map. The volumetric caution and warning cueing regions fed by the radar hits are referred to as a Risk Space. The Risk Space is speed dependent and parametric, primarily defined by caution and warning time-to-collision values (Ref. [20]). Cueing outside of the Risk Space volume is considered advisory (see Figure 5). A 2D risk space overlays the plan-view moving map, as seen in Figure 6, Left, with a caution region (yellow contour line) and a warning region (red contour line). Matching colored symbology is used for the depiction of the obstacles on the PFD (see Figure 6, Right). Layered cueing is employed such that an additional sensory modality is introduced as a hazard transition from an advisory to a caution to a warning. When entering the caution region of the risk space, the two most-urgent obstacles/ terrain are depicted by auditory icons presented spatially via headphones (i.e., distance and position relative to the listener). When entering the warning region, supplementary tactile sensory cues are made available via a tactile situation awareness device (see annex for details). In a series of experiments, in simulator and in flight, ICE-CAS proved to both reduce WL and increase SA [Ref. [19]].

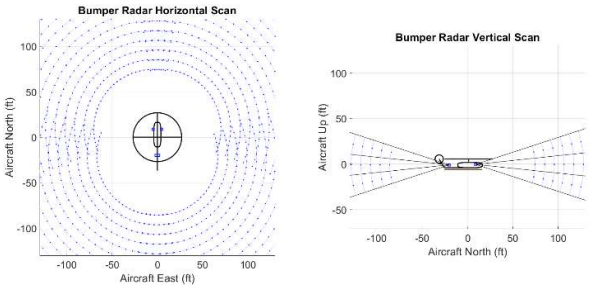


Figure 4. Simulated Radar Scan. Radars scan 360° azimuth (4° increment) at three elevations: -12 °, 0°, +12°.

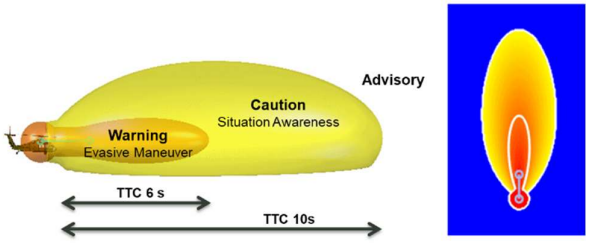


Figure 5. Left. Risk Space: Speed-dependent caution and warning cueing regions defined by caution and warning time-to-collision (TTC) values. Right. Risk map: Ranks radar hits in terms of urgency. The two most urgent obstacles are cued.

Figure 6. ICE-CAS Visual Symbology. Left: Multi-Function Display (MFD). Right: Primary Flight Display (PFD). Top: The ICE-CAS radar-hit obstacle symbology. Bottom: The ICE-CAS power-line symbology with bounding-box dimensions based on the Federal Aviation Agency (FAA) obstacle database positional uncertainty.

Flight Control (FC) LOA: MAA

The Mission Adaptive Autonomy (MAA) is a set of guidance and flight control software that provides a range of behaviors ranging from stabilized Attitude Command Attitude Hold (ACAH) response that a pilot can fly manually to fully reactive autonomy that the pilot can interact with low WL at an executive level. The system has a range sensor to "see and avoid" obstacles as it works its way to a pilot selectable destination. Once at the destination it scans the landing zone (LZ) and finds safe areas on which to land (see Figure 7). Combined with the ICE display (Ref. [21]), MAA provides provide sufficient situational awareness for the autonomy to fly the mission in DVE while reducing the WL. The two levels of autonomy, and the two levels of obstacle cueing are between route factors, allowing full factorial data analysis, i.e., each pilot experiences all four conditions.

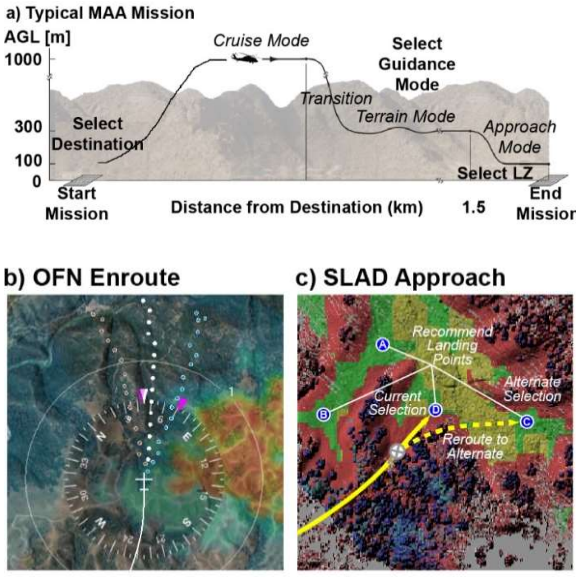


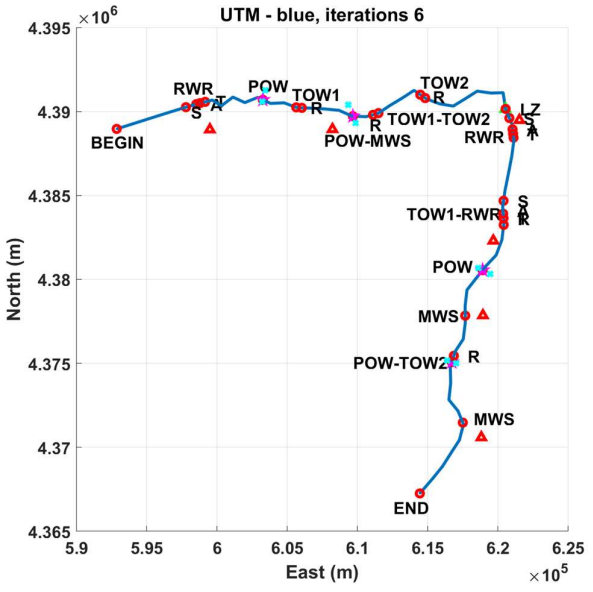
Figure 7: Elements of Mission Adaptive Autonomy (MAA): a) Typical *MAA* Mission Manager, b) Obstacle Field Navigation (OFN) enroute (Cruise or Terrain mode) with bias guidance by heading or glide slope, and c) Safe Landing Area Determination (SLAD) approach with Landing Zone *(LZ)* selection.

b. The Local Factors

For the local factors, five hypothetical levels of Workload (Very Low, Low, Medium, High, Very High) were be manipulated within each route. The different levels of workload were controlled by the type of events encountered during the flight such as obstacles (further referred to as Route Obstacles), hostile fire (further referred to as ASE events), and the mission phase (enroute, approach to hover, hover at pick-up point, take-off from hover). Each of the five routes consisted of geographically fixed and scripted elements (see Figures 8 and 9). The fixed elements included the starting location, a MEDEVAC pick-up point, four power line crossings, and an ending location. The scripted elements included obstacle popup triggers and Aircraft Survivability Equipment (ASE) threat events. To ensure that the pilots would experience a wide range of WL, the events 1) were selected to represent different levels of WL and 2) were presented in isolation or combined, as seen in Table 3. Each route included the following scripted events: one tower1 alone, one tower2 alone, two RWRs, two MWSs, and one of each combination. Thus, three tower1, tower2, RWR, and MWS events occurred alone or in a combination per route. Along with the geographically fixed events (power line crossings and medevac), this was considered an appropriate level of route activity as judged by pilots during the design phase. The event sequence and corresponding time intervals were randomized, when possible, with the power lines and the medevac placing limitations on event placement. At this point, there is no assumption that Obstacle events and ASE events are associated with different levels of WL. Meanwhile, powerline crossing is expected to generate a higher level of WL than either the Obstacle or ASE events. Last, the combination of events should be associated to the highest level of WL, in particular for the combination of a powerline and an event.

Table 3. Predicted WL as a function of Events and Phases of Flight (Additive Model).

Events		Hover	Take-Off	Enroute	Approach1	Approach2	Approach3 *=no event
	WL	+2	+2	+1	+2	+3	+4*
No Event	0	2					4*
RWR Launch	1		3				4*
MWS Launch	1			2			4*
TOW1	1				3		4*
TOW2	2					5	4*
TOW1-RWR	2						4*
POW (above clearance)	2					5	4*
TOW1-TOW2	3						4*
POW (below clearance)	3						4*
POW-MWS (above clearance)	3						4*
POWS-TOW2 (above clearance)	3						4*
POW-MWS (below clearance)	4						4*
POWS-TOW2 (below clearance)	4						4*



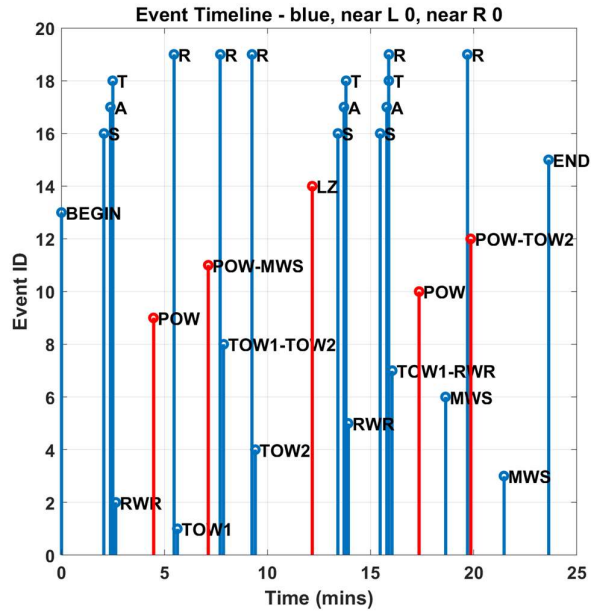


Figure 8: The Blue route (experimental trial). Estimated duration 23:38 mins. Top: Distribution of the Obstacles and ASE events superimposed on the trajectory. Red circle: route event; red triangle: ASE location where the spatial location is based on the ASE launch event; cyan x: power line towers, green triangle: MEDEVAC Landing Zone (LZ); blue dot: ground track vector end point; S: RWR Search; A: RWR Acquisition; T: RWR Track; POW: Powerline; TOW: Tower; R: tower popup trigger for the next tower event (i.e., 500 m (1640 ft) before tower). Bottom: Obstacles and ASE events timeline.

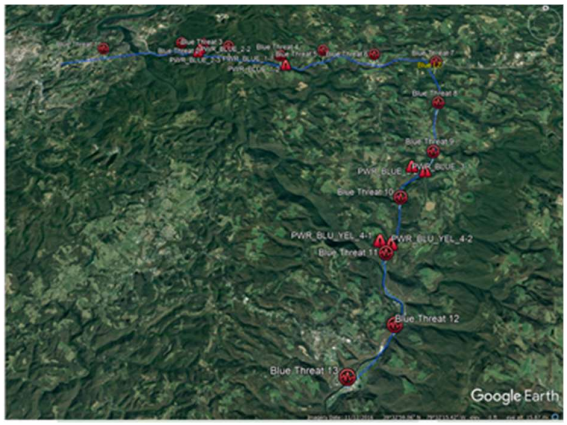


Figure 9: Scripted events distribution during an experimental route (Blue route).

c. Mission Task Elements (MTEs): Gradual changes

Each of the route was structured in an identical way regarding the mission phase. The mission started at a 150-ft hover, followed by an enroute phase, an approach to hover approximately in the middle of the route, followed by a 1-minute hover at 75 ft to simulate a patient pick-up, take-off from hover, a second enroute phase followed by as second approach to hover with an ending at a 75-ft hover. Each mission phase was assumed to be representative of a different level of workload, as seen in Table 3 and Figure 10. Enroute when no events were present was expected to be associated to the lowest level of workload, i.e., Low. Note that the baseline for workload was set at Very Low (1) at the start of the trial. The Approach to hover phase was decomposed into three successive events: 1) distance <= 1 NM (WL=Medium,), 2) glide slope indicator on (WL=High), and 3) altitude AGL<=100 ft and /or GS<=25 kts (WL= High).

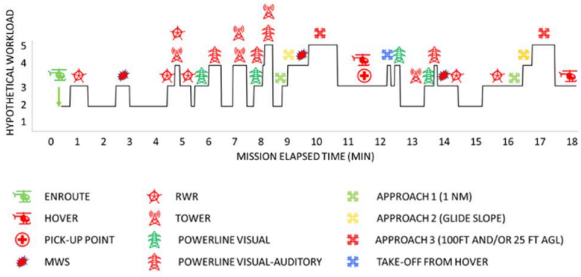


Figure 10: Times series for the Hypothetical Workload associated to the different local events for the White route (training) and Phases of the flight. The events are randomized between conditions routes and between pilots. The effect of Terrain is not included in this representation.

Once the hover was stabilized, the expected workload level was set at Medium. Take-off from hover predicted workload level was set at High. It was also expected that the route difficulty, in terms of terrain profile requiring sudden changes in speed and altitude, would also contribute to the perceived workload.

Hypothetical WL

In an effort vary pilot WL, various obstacle and ASE events were scripted to occur at random locations along the route and, hence, during different phases of flight and with varying terrain navigation difficulty. Thus, the designed workload-inducing events ultimately consisted of three components, the event, the phase of flight, and the terrain. This resulted in an additive model in regard to a pilot-independent hypothetical (predicted) WL value. As a first iteration, integer values were associated a posteriori with the obstacle/ASE events and the phase of flight as shown in Table 3. The terrain gradient was used to capture the terrain variation and, hence, anticipated navigation difficulty and pilot WL.

Given the terrain varied with the specific route flown by the pilot, the simulated-EGI MSL altitude and radar altitude values were differenced to obtain the terrain altitude profile (Figure 11). The derivative of this time-domain sequence yielded the terrain gradient with respect to time of flight. Since both positive and negative deviations were considered equally difficult, the absolute value of the gradient was used. Given WL was polled every 30 seconds, a 30 second moving average filter was applied in an effort to capture the integrative effect of the pilot's memory for the preceding polling window. If this were the only effect modeled, the filter's lag would have been left uncompensated at 15 seconds. However, the route ¼-mile visibility conditions resulted in the pilot being able to evaluate the demands of the next 15.6 seconds of travel (when ground speed equal to the target of 100 knots).

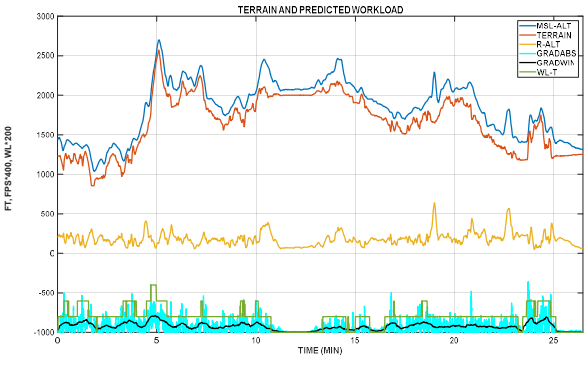


Figure 11. Altitude and terrain gradient variables used to estimate the terrain contribution to the Hypothetical Workload value.

Thus, the filter lag was nullified to adjust for this anticipatory effect. From observation, a scalar was determined that would result in a maximum rounded terrain effect value of 3, on average.

The preceding steps yielded hypothetical WL values for three components, event, phase of flight, and terrain, at every 60 Hz simulation update rate sample of time. The three components were then summed and clamped to a maximum value of 5 to serve as a hypothetical workload basis of reference for the participant's workload entries.

2.1.2. The Simulation Environment

a. The simulator

Two simulators one located at Moffett Field, CA and one at Ft. Eustis, VA have been equipped for this research. The Moffett Field System Integration Laboratory (SIL) simulator, shown in figure 12, Left, was used for software development and initial scenario development.

Figure 12: Left: System Integration Laboratory (SIL) at the NASA Ames Research Center (NASA ARC) Moffett Field, CA. Right: rotorcratf in flight (RIFL) SIL at Ft Eustis, VA.

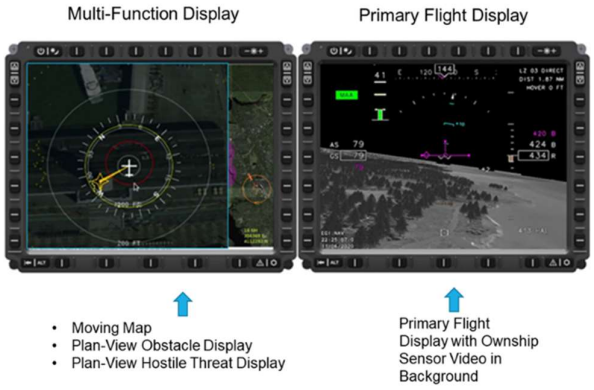


Figure 13. Panel-Mounted Displays for both SIL simulators. Left: Multi-Function Display (MFD) with moving map, plan-view obstacle display and plan-view hostile threat display. Right: Primary Flight Display (PFD) with Ownship infra-red (IR) sensor video in background.

Figure 14: Left: Participant wearing the Sennheiser Game 1 headphones with head-tracker, EEG sensors and Pupil Lab Core eye-tracker. Right: Tactile Seat Bottom Cushion (Engineering Acoustics Inc.)

The Ft. Eustis rotorcraft in flight (RIFL) SIL simulator, shown in figure 12 right, was used for scenario refinement and the conduct of the actual data runs. The two simulators are nearly identical. They are both equipped with a UH-60 instrument panel with two 10-inch diagonal panel mounted displays (PMDs, see Figure 13). Both SIL simulators have a large monitor(s) for the out-the-window (OTW) view, same electrically back-driven control sticks and pedals, identical computers driving the helicopter model, OTW scene generator, Primary Mounted Display (PMD) generator (ICE), Aviation Auditory Display Engine (AvADE), and the autonomous flight control system (Mission Adaptive Autonomy - MAA). Therefore, software can be ported directly between simulators. In the future (not part of this test plan), software can also be ported directly the EH-60L research aircraft tail number 87-24657, since the SILs were built to support the aircraft.

b. Data recording

The Fort Eustis SIL was temporarily equipped with the biometric measurement equipment listed in Table 4 and illustrated in Figure 14 (ECG sensors not shown).

Electroencephalography. EEG data were recorded using the Advanced Brain Monitoring (ABM) B-Alert X-24® wireless wet electrode system (256 Hz sampling rate) with 20 electrodes placed according to the International 10-20 system. Frontal sites examined included: Fp1, Fp2, F7, F3, Fz, F4, and F8. The EEG signal was referenced to linked mastoid electrodes. A 0.5 Hz – 65 Hz band pass filter was applied to the signal. Notch filters (50, 60, 100, 120 Hz) were used to sharply attenuate environmental noise not reduced by the bandpass filter. PSD values were computed using the automated algorithms provided through the B-Alert Live® data acquisition software (ABM, 2009). Prior to computing PSD values, artifacts (spikes, excursions, amplifier saturations, electromyography, eye blinks) were identified and removed using algorithms developed by ABM (ABM, 2009; Ref. [22]). The software Fast Fourier Transformed (FFT) the raw EEG signal data and calculated the amplitudes of the sinusoidal components for designated frequency bins. Frequency domain variables were based on the PSD derived after application of a 50% overlapping window and Kaiser window. Each window size was defined as one epoch containing 1 s of data (256 decontaminated samples). The software then provided PSD values ranging from 1 to 40 Hz for each EEG channel that were logged to obtain a Gaussian distribution. Selected relative 1 Hz bins were averaged to create the EEG bands used in analyses (theta: 4-8 Hz; alpha: 9-13 Hz; beta: 14-30 Hz). PSD values were examined from frontal channels with values averaged across each task. An engagement index of the ratio of beta values to alpha plus theta values (Ref. [23]) was also examined.

Electrocardiogram (ECG). The Biopac BioNomadix electrocardiogram amplifier module (ECG100) collected electrocardiogram (ECG) data. ECG activity with the Biopac BioNomadix records through single-lead electrodes placed on each of a participant's clavicles and one below the right pectoral area. Data samples at a rate of 1,000 hertz (Hz). The single-lead electrodes connect to a wireless transmitter belt worn around the torso that transmits the signals to a computer for recording and later processing. The research team recorded participant baseline data after ECG electrode placement. Baseline data recording required the participant to sit in a relaxed position for 5 minutes. During the baseline recordings, a member of the research team monitored the signal to ensure a clean signal is collected, indicating accurate placement of electrodes. New baseline recordings were made with each new placement of the ECG electrodes. Data were processed (including filtered, transformed, and artifacts removed) using the peak detection algorithm described in Ref. [24].

Table 4. Biometric Measurements

Electrocardiogram (ECG)	Heart activity
Electroencephalogram (EEG)	Brain activity
Eye and head tracking	Measure of the direction of the head (3D) and direction of stare of the eyes (2D, 3D)
Pupillometry	Measure of the eye pupil dilation

This algorithm was used to identify R-wave peaks and identify the inter-beat-intervals (IBI); from the IBI, the outcome measures were extracted. Outcome measures included heart rate, measured in beats per minute, and heart rate variability, measured using the standard deviation of normal-to-normal (SDNN) beats.

Eye-tracking and pupillometry. Eye tracking is the process of measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head. An eye tracker is a device for measuring eye positions, pupil dilation and eye movement. Pupil Core is an eye tracking suite that is comprised of an open-source software suite and a wearable binocular eye tracking headset. It was selected for the headset's light weight and modularity (see Appendix C for details). Available metrics include blink detection, fixation and saccade detection, pupil detection (2D, 3D), gaze position, surface (areas of interest, AOIs) tracker**.** In addition to these equipment, audio and video recording was performed for a posteriori analyses.

c. The participants

Fourteen Active Duty or Department of the Army Civilian Experimental Test Pilots from the Rotorcraft In Flight Laboratory (RIFL) at Fort Eustis were recruited to participate in the experiment. Age ranged between 32 and 60 years (mean: 41.38), flight hours between 1300 and 7000 (mean: 2758).

d. The procedure

Participants were equipped with headphones, a tactile situational awareness system (TSAS) vibration cueing system consisting of a bottom seat cushion as well as ECG, EEG and eye-tracking glasses. Each pilot spend a full day at the simulation facility. The morning was dedicated to the training, which included briefing and familiarization with the simulator and the task. The experimental sessions took place in the afternoon after a lunch break. They were followed by the administration of a post run questionnaire and subjective rating scales, a post-experiment questionnaires and a debrief.

e. The Task

The pilots were be requested to conduct a tactically relevant MEDEVAC mission in four different conditions (2 levels of Autonomy x 2 levels of Cueing), each condition being associated to a different route to control for the effects of terrain. The evaluation pilot alternated between manually flying the aircraft (two routes) or monitoring the autonomous flight control system and occasionally puting in pilot-on-the-loop inputs to direct the flight control system to change the flight path left/right or higher/lower (two routes). The command model was attitude command.

In each condition, Manual mode vs. Autonomous mode, the two types of symbology were tested.
In the Nominal condition, display symbology included the Integrated Cueing Environmenmt (ICE) symbology set. The visual cueing includes an image of the terrain, artificial objects geo-referenced to that terrain (such as an artificial landing pad and waypoint markers), aircraft state information, flight control mode status, and guidance cueing for en route, approach-to-landing, and approach-to-hover. A monaural voice synthesizer provided advisories, cautions, and warnings unique to the Degraded Visual Environment- Mitigation (DVE-M) system. A tactile seat cushion cues the pilot for excessive vertical speed.
In the Obstacle Cueing condition, the ICE symbology was complemented with the multimodal visual-auditory-tactile symbology ICE-Collision Assistance System (ICE-CAS).
For each route, the in-bound leg was 10-15 minutes At the pick-up point the aircraft had to come to a hover, maintain hover for approximately 1 minutes for a simulated hoist pick-up of the patient, and then take-off from hover. The outbound leg was also be 10-15 minutes long, and ended with a hover at a different destination, briefed as a hospital.
Pilots were instructed to fly the MEDEVAC mission at an altitude Above Ground Level (AGL) of 150ft or lower (+15ft max) and at a speed of 80 kts (+/- 10 kts) for the enroute phase, and to follow the commanded altitude and speed when in approach and hover (75ft). A magenta ground track was overlayed on the MFD moving map (see Figure 15, Left) and associated magenta chevron were superimposed on the IR image on the PFD (see Figure 15, Right).
A research team member was seated next to the pilot to provide feedback regarding altitude and speed in case of prolongued excursions from the prescribed range. Pilots were taught to fly over the powerlines and to manoeuver around the obstacles as a function of their affordances (go left or right, fly over). Pilots were told ahead of time of the types of events that will occur, but not where or when they will occur. Pilots took a minimum 10 minutes break between each route. In parallel to the flying task, pilots were instructed to report their perceived WL level used a knob on the cyclic to adjust the orange integer representing current workload on their out-the-window (OTW) display (see Figure 15, Center and Right). An audio signal prompted the pilot for a WL update every 30 seconds. If the response was not provided within 5 seconds, the integer would flash for the next 5 seconds, then get back to its nominal configuration (not flashing). Pilots were also told to update their perceived WL as they felt it was changing, between prompts to provide a finer grain for the analysis of the WL triggers. If the WL was updated between prompts, the next prompt occurred 30 seconds later.

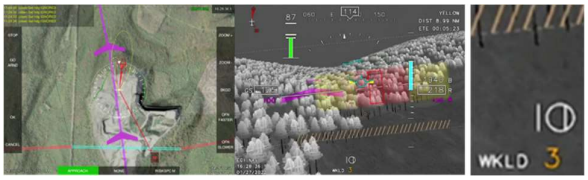


Figure 15. Left: Multi-Function Display (MFD). Center: Primary Flight Display (PFD). Right: Detail of PFD with displayed WL level. Magenta ground track was overlayed on the MFD moving map, associated magenta chevron were superimposed on the IR image on the PFD.

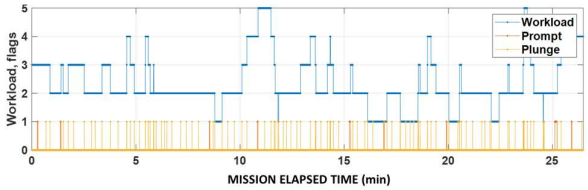


Figure 16. Times series for for a training trial (Yellow route, Pilot p04) in the Cueing ON, MAA uncoupled condition for WL, prompts and recorded action on the cyclic for WL update (plunge).

Figure 16 depicts the time series for a training trial (White route) in the Cueing ON, MAA uncoupled condition for WL, prompts and recorded action on the cyclic for WL update (plunge). At the end of each route, pilots were asked to report their WL with the NASA-TLX mobile app developed by NASA ARC (Ref. [25]).

f. The measures of WL

Objective Measures

Control Attack

Padfield (Ref. [26]) formulated the control attack metric, A_η, based on the rate and magnitude of the pilot's control input to quantify agility. Control attack characterizes each discrete control input and is defined as the ratio of the peak rate of control displacement, $\dot{\eta}_{pk}$, to the magnitude of the change in the control displacement, $\Delta\eta$,

$$A_\eta = \frac{\dot{\eta}_{pk}}{\Delta\eta}$$

A higher attack value indicates rapid-small control deflections while a lower attack value indicates slower-large control deflections. An example of a high attack and low attack pilot control input are shown in **Figure**7.

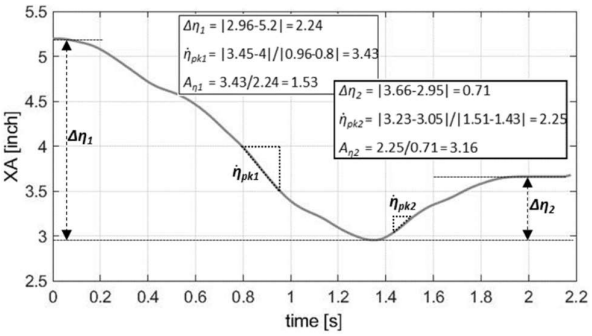


Figure 15. Control deflection time history showing high and low values of attack (Ref.[27]).

Each axis of control signal is filtered at 3 Hz, and local maxima and minima (max, min) within a 10-second sliding window are identified. The maximum absolute control rate associated with two successive max/min or min/max pairs (and the corresponding control deflection) is used to compute the control attack for that window in time. For each time window A_η for each of the four control axes (cyclic pitch and roll, collective, and pedals) are added together, producing an aggregate control attack value.

The Dynamic Interface Modeling and Simulation System Product Metric (DIMSS-PM)

The Dynamic Interface Modeling and Simulation System Product Metric (DIMSS-PM, Ref. [28]) is a pilot workload metric that was used to evaluate helicopter shipboard launch and recovery. The basic principle underlying the DIMSS-PM is that high pilot workload, represented by large DIMSS values, arise from high frequency, large amplitude control movements. Conversely, low pilot workload is represented by low DIMSS PM values that result from low frequency and low amplitude control movements. The metric is composed of the product of the number of control reversals and the standard deviation of control deflections in a moving window (control motion is filtered at 3.3 Hz). It was calculated on a sliding window $[Wn; Wn+1]$ as follow:

$$DIMSS[w_n;\ w_{n+1}] = (Nb_{maxima} + Nb_{minima}) * Std(stick[w_n;\ w_{n+1}])$$

where $(Nb_{maxima} + Nb_{minima})$ is the number of control reversals with the stick multiplied by the standard deviation (Std) of the stick value; Nb_{maxima} and Nb_{minima} being respectively the number of local maxima and the number of local minima on the stick value curve. The higher the value, the less stable the joystick control. A zero value means that pilots did not move the joystick.

Each axis of control signal is filtered at 3 Hz, and the total number of local maxima and minima (max, min) within a 10-second sliding window are identified. This number is multiplied by the standard deviation of the

control deflection to produce the DIMSS metric. For each time window, DIMSS for each of the four control axes (cyclic pitch and roll, collective, and pedals) are added together, producing an aggregate DIMSS value.

Spare Capacity OPerations Estimator (SCOPE)

In Ref. [29] Bachelder showed how the basic laws of sensory perception (laws of Ekman, Weber and Stevens, Refs. [30]) can be extended to multi-input sensing, where that multiple sensory inputs add logarithmically. Or equivalently, multiple sensory inputs multiply. Bachelder proposed that if a perceptual response R is composed of multiple stimuli Si each having its constant of proportionality γi such that

$$\frac{\Delta R}{R} = \sum_{i=1}^{N} \gamma_i \frac{\Delta S_i}{S_i} \Longrightarrow \ln R = \sum_{i=1}^{N} \gamma_i \ln S_i + \ln \lambda \Longrightarrow R = \lambda \prod_{i=1}^{N} S_i^{\gamma_i}$$

γi represents the perceptual exponent associated with each stimulus Si, and the technique employing the product of exponented stimuli is referred to as the Spare Capacity OPerations Estimator (SCOPE, Ref. [31]).

For compensatory tracking tasks, Bachelder observed (Ref. [32]) that the stimuli for workload perception were well represented by the standard deviation of the rates of the cues employed by the pilot for control (i.e., error and control), and the perceptual exponent for each cue was 0.07. It is shown that this value corresponds to the optimal balance between information value (ability to resolve stability) and linearity (which affects the ease that workload perception can be transformed into a metric of stability).

In Ref. [33] SCOPE was applied to the highly coupled and visually complex task of aggressive approach-to-land, producing high correlation with actual workload collected in real time. In addition to control-related variables, cues contributing to pilot workload for this task also reflected operational/usability constraints (such as specific kinetic energy and approach angle permitting visibility of the landing zone). The perceptual exponents of cues associated with control were 0.07, but for the other cues the exponent varied from 0.025 to 0.14, where a higher value indicates a greater influence on workload perception.

The stimuli that SCOPE incorporates are not confined to a particular sensing modality – in Ref. [34] SCOPE employs vision and proprioception (visual error, limb force, and limb position). Thus, any cue impinging on the operator's sensorium having a significant effect on workload perception can be used by SCOPE. In Ref. 29 it is emphasized that the multiplicative nature of SCOPE frees it from the difficulty associated with mixing inputs of different scaling and units – these effects are all subsumed by a single constant of proportionality that is multiplied with the product of the exponented stimuli (the variable λ in the above equation).This is analogous to behavior observed in biological circuits that exhibit a property called fold-change detection, described in Ref. [35]. When these circuits detect multiple types of inputs (stimuli), the integrated response is multiplicative. Under certain conditions the stimuli are weighted differently through exponentiation, and then multiplied. This process allows for mixed input types having different scaling.

The suite of variables and their exponent weightings that produced the highest correlation with the recorded SIMEVAL workload are given below. The standard deviation of each variable using a sliding window of ten seconds was employed to compute the workload estimate. A variable's exponent allows comparison of that variable's influence workload. Collective control rate ($\dot{\delta}_z$) and vertical sink rate ($[V_z$) are seen to have four times the effect on workload compared to control of the longitudinal and yaw axes. Lateral axis control had the least effect on workload (16 times less than vertical control). The constants *C* and *D* are multiplicative and additive constants, respectively.

$$WL_{SCOPE} = C\left\{\Pi\left([V_y\dot{\delta}_y]^{0.25}, [V_x\dot{\delta}_x], [\dot{\Psi}\dot{\delta}_\Psi], [V_z\dot{\delta}_z]^4\right)\right\}^{0.04} + D$$

With:
V_x: $longitudinal\ speed$,
V_y: $lateral\ speed$,
V_z: $vertical\ speed$,
$\dot{\Psi}$: $yaw\ rate$,
$\dot{\delta}_x$: $longitudinal\ cyclic\ rate$,
$\dot{\delta}_y$: $lateral\ cyclic\ rate$,
$\dot{\delta}_z$: $collective\ rate$,
$\dot{\delta}_\Psi$: $pedal\ rate$,
C: $multiplicative\ constant$,
D: $bias\ constant$

f. Methodology for data processing and Analysis

As specified previously, the experimental routes were designed to create "test points" to increase locally the WL at different levels for different conditions of global workload. In terms of data processing and data analysis, the main goals of the simulation are:

- To express the continual workload as time series (data that have been observed at different points in time), giving rise to multiple workload profiles obtained via the different metrics (reported WL, EEG, ECG, eye-tracking, control inputs). The systematic collection of the WL rating at every 30 sec or less offers the possibility to correlate the subjective data with the qualitative measures of performance.
- To identify and quantify the relationship between two or more variables by using multiple statistical

methods: Correlation analysis, principal component analysis (PCA), confirmatory factor analysis (CFA), structural equation model (SEM) and Bayesian network approach.
- To compare the global effects of WL using a fully factorial design and factorial analysis: repeated-measures analysis of variance (ANOVAs).
- To develop a model from which performance can be continuously estimated from WL.

3. RESULTS

The results presented here are limited to the Manual flight condition only (MAA uncoupled).

Table 5 represents the 5-point scale used by the pilots to report their perceived WL.

Table 5. On-line WL report on a 5-point scale.

Real-time Rating Scale	1	2	3	4	5
Associated WL level	Very Low	Low	Intermediate	High	Very High

3.1. Reported and Hypothetical WL

3.1.1. Reported WL

Pilot p13 was excluded from the analyses because of missing data in the manual cueing off condition.
First, there was no significant effect of route on the reported WL ($F_{(1,22)} = .03, p = .95$), meaning that the overall difficulty associated to the route profile was not a factor in the results.
Overall, mean WL was intermediate (μ =2.97, SE=.03, min=1, max=5). Univariate Analyses of Variance (ANOVAs) were performed to assess the effect of cueing condition (ON, OFF) and pilot # on the distribution of WL between ratings. The effect of condition was not significant ($F_{(1,24)} = .007, p = .93$), but one can see from Figure 16 that there was a greater intra-individual variability between pilots in the Cueing ON condition. The effect of pilot was highly significant: $F_{(1,13)} = 26.81, p < .0001$). Pilots p02, p05 and p14 reported a significantly higher WL than average, while pilots p04, p11 and p12 reported a significantly lower WL than average. It is interesting at this point to look at the distribution of the reported workload over time, in terms of amplitude and frequency of updates.

Table 6. WL distribution. Multivariate effects for Condition and Pilot

Variables	Pillai's Trace	*F*	*df*	Error *df*	*p*
Condition	.005	.02	5	20	1
Pilot	4.32	52.89	60	65	<.001

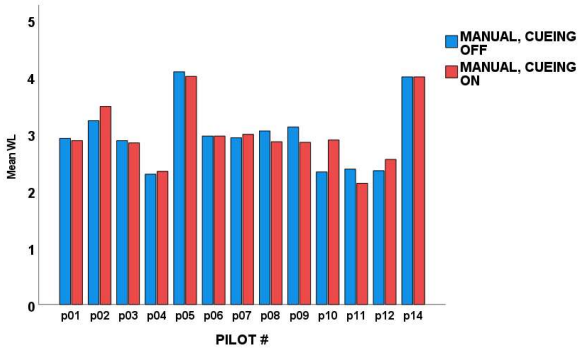


Figure 16. Mean reported WL level for the 13 pilots in the Manual condition as a function of Cueing, OFF, or ON.

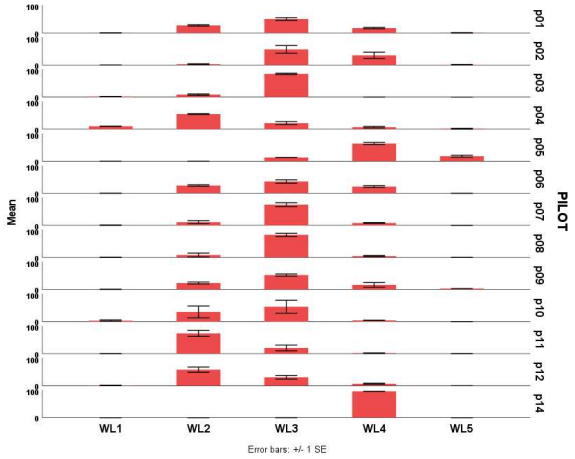


Figure 17. Distribution of the mean reported WL ratings for the 13 pilots.

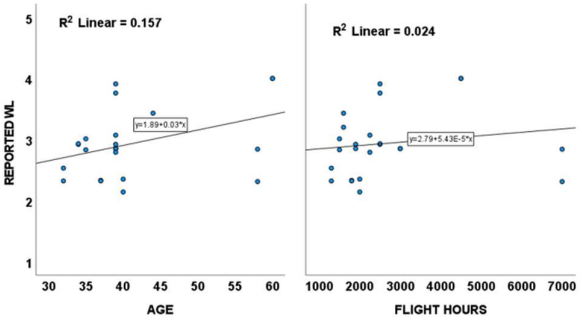


Figure 18. Linear regression for Reported WL as a function of pilot's age (Left) and flight hours (Right).

Pilots were prompted to report their WL every 30 sec. They could also modify their ratings within this time window if they felt their WL level had changed. Figure 17 illustrates the distribution of the rating for each pilot.

A multivariate ANOVA showed that there was no significant effect of the condition on the distribution of ratings. Conversely, the distribution of ratings significantly varied among pilots (see table 6).
A multiple regression (method Enter) was performed to explore the effects of age and experience on the Reported WL (see Figure 18). The combined effects of Age and Flight Hours was statistically significant, essentially the resultant of the effect of age ($R^2 = .26, p = .03$; Flight Hours: $Standardized\ \beta = -.61, t = -1.67, p = .09$; Age: $Standardized\ \beta = .91, t = 2.63, p = .01$).

3.1.2. Frequency of Plunges

The frequency of plunges (WL rating updates) was not significantly different between cueing conditions, ON vs. OFF ($F_{(1,24)} = .000, p = .98$). . The number of changes varied from 0 to 49 in the Manual, Cueing OFF condition and from 1 to 42 in the Manual, Cueing ON condition. Conversely, pilots' frequency of plunges was significantly different between pilots $F_{(1,13)} = 26.81, p < .0001$), as seen in Figure 18. The frequency of plunges was inversely correlated to Age ($R^2 = -.52, p = .009$) and to a less extent, inversely correlated to Flight Hours ($R^2 = -.47, p = .01$).
Figure 19 shows the number of plunges and the reported WL for a run in the Manual, Cueing ON condition for two representative pilots, p04 for low WL, and Pilot 5 for high WL. Pilots were categorized into three groups, Low (p04, p11, p12) Intermediate (p01, p02, p06, p07, p08, p09, p14) and High (p03, p05, p14) based on the level of WL. A correlation analysis was performed to assess the degree of association between the mean value of the reported WL and the number of changes made by the pilots. A negative correlation of - .46 (N=26, *p*=.01) was obtained, highlighting the fact that higher mean levels of WL were associated with lower number of WL changes, as seen in Figure 20). This can be interpreted as a consequence of the level of self-awareness, or and or overall efficiency level.

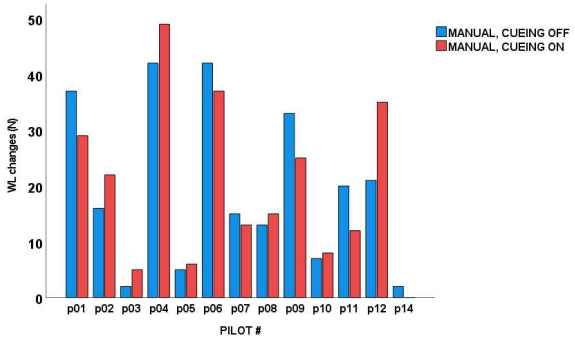


Figure 18. Number of WL level changes for the 13 pilots as a function of Cueing type. Note 0 changes for p14 in the Manual, Cueing ON condition.

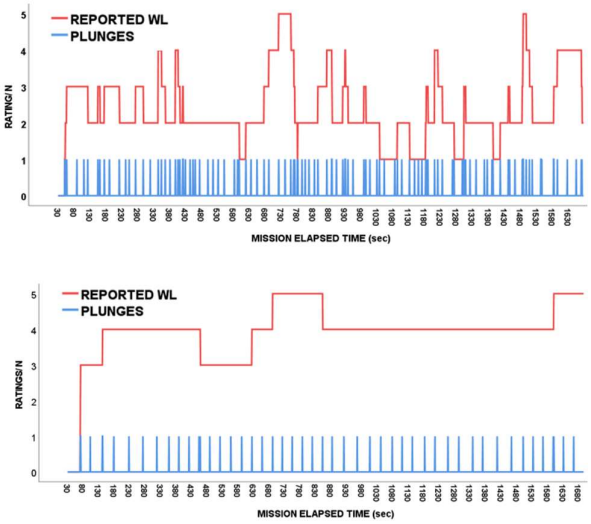


Figure 19. Top: Pilot p04 (Reported WL: $\mu = 2.36, SD = .91$; N Plunges = 118), Bottom: Pilot p05 (Reported WL: $\mu = 4.01, SD = .55$; N Plunges = 50).

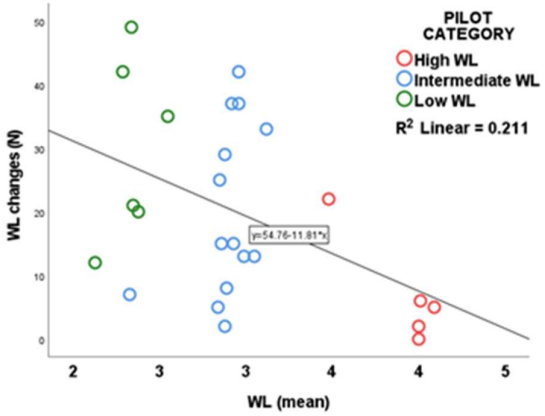


Figure 20. Regression plot for the number of Plunges as a function of the mean reported WL ($R^2 = .21, p = .01$).

Most self-aware/ efficient pilots had a wider range of reported levels of WL. Difference was significant only between Low and High. These results will need to be analyzed in relation with performance metrics such as deviation from speed, altitude, and heading, and with control activity.

3.1.3. Hypothetical WL: External Factors

The mean hypothetical workload for each pilot and each condition is illustrated in figure 21. Because the hypothetical WL algorithm does not take into account the Cueing type, there was no significant difference in the hypothetical WL between cueing conditions ($F_{(1,24)} = 1.53, p = .22$). For the same reason, there was no significant effect of pilot ($F_{(1,13)} = .76, p = .67$).

a. Correlations between reported and hypothetical WL

Two examples of the co-variations between Reported and hypothetical WL are presented in Figures 22 and 23, for a pilot representative of the Low WL, high frequency of WL level changes (p04) and a pilot representative of the High WL, and low frequency of WL level changes (p05). The contribution of each dimension of the Hypothetical WL, Event, Terrain, and Phase was assessed.

For Pilot p04, Manual, Cueing On condition, the correlation between Reported WL and Hypothetical WL was highly significant ($R^2 = .36, p < .0001$). The most significant correlation was observed with Phase WL ($R^2 = .53, p < .0001$), then in decreasing order, with Terrain WL ($R^2 = -.103, p < .001$) and Event WL ($R^2 = .08, p < .001$). The relatively modest correlation between WL and Event WL and WL and Terrain could be explained by the fact that pilots may have been flying higher than instructed, or if the heading was not maintained enough so that the obstacle events wouldn't be encountered. For Pilot p05, the correlation between Reported WL and Hypothetical WL was lower, as expected from the difference in temporal resolution. We observed similar trends regarding the Hypothetical WL components. The most significant correlation was observed with Phase WL ($R^2 = .31, p < .0001$), then in decreasing order, with Terrain WL ($R^2 = -.06, p < .001$) and Event WL ($R^2 = -.05, p < .001$).

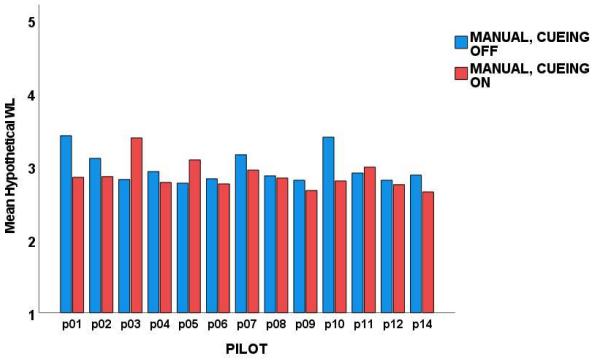


Figure 21. Mean Hypothetical WL level for the 13 pilots as a function of Cueing condition: OFF, ON.

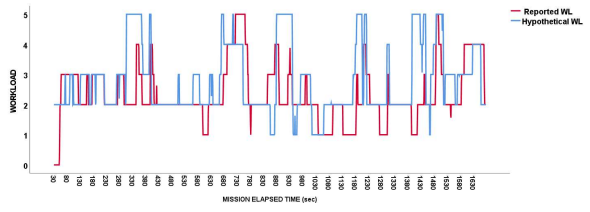


Figure 22. Time series for Reported and Hypothetical WL for Pilot p04, Manual, Cueing ON condition. R2=.35, p<.0001

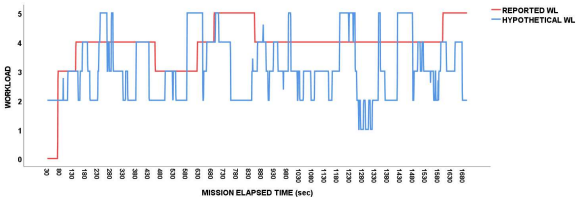


Figure 23. Time series for Reported and Hypothetical WL for Pilot p05, Manual, Cueing ON condition. R2=.16.

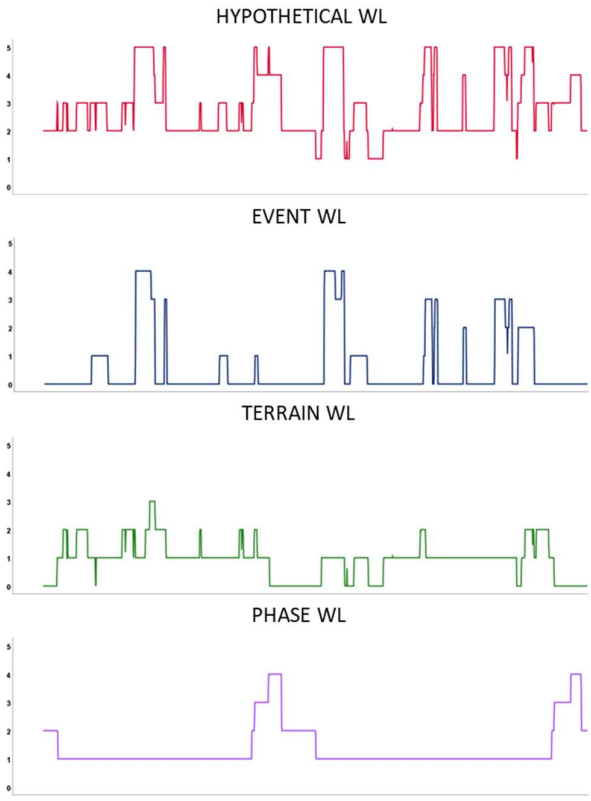


Figure 24. The 3 components were summed and clamped to a max value of 5 to serve as a Hypothetical WL reference for the participant's WL entries. Pilot p04, Cueing ON condition.

A factorial analysis was performed on the data for all pilots to assess the relationships between the three components of the hypothetical WL.

The only significant correlation was between Terrain and Phase ($R^2 = -.17, p = .04$). The Bartlett's test of sphericity was not significant ($X_3^2 = 4.67, p = .19$). Initial eigen values indicated that the first two factors (components) explained 44%, and 37% of the variance respectively. The solution for the two factors were each examined using a varimax rotation of the factor loading matrix. Component 1 included Phase and Terrain, Component 2, Events (see Table 7).

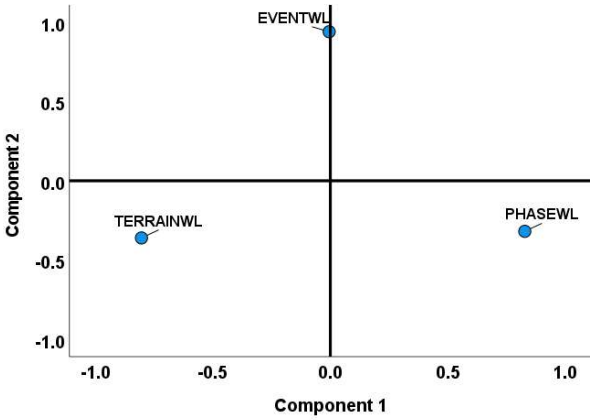


Figure 24. Component Plot in Rotated Space for the Hypothetical WL. Extraction Method: PCA. Rotation Method: Varimax with Kaiser Normalization

For Component 1, Terrain and Phase were negatively correlated ($R^2 = -.33, p = .04$), as illustrated in Figure 24. The Terrain and Phase parameters could then be combined in the next iteration of Hypothetical WL computation.

Last, it was observed that the Hypothetical WL was strongly correlated to Terrain ($R^2 = .87, p < .001$). Correlations with Phase and Events were not significant. The lack of correlation with events could potentially be explained by the fact that pilots could have been flying either too high or deviated enough from the commanded heading not to be sufficiently exposed to the obstacle events. Requesting pilots to adopt a specific strategy to negotiate the obstacles could potentially increase the contribution of Events to the WL. Also, the relative contribution of ASE events and Obstacle Events to the hypothetical WL remains to be assessed.

Table 7. Component Score Coefficient Matrix for the Hypothetical WL components.

	Component	
	1	2
EVENT WL	-.017	.842
PHASE WL	.626	-.298
TERRAIN WL	-.598	-.311

3.2. SCOPE, DIMSS and Control Attack Metrics: Internal Factors

3.2.1. Spare Capacity Operator Estimate (SCOPE)

Preliminary analyses validated the WL driver's selection, and the usability of the real-time subjective WL report using a real-time rating scale, used to compute SCOPE. SCOPE was applied to a 20-minute segment of flight data collected from one pilot in the training route. For this initial analysis only the rates of the variables associated with manual flight control were employed (i.e., aircraft motion and the rates of their primary controls).

Overall, the correlation between Reported WL and SCOPE was relatively high with $R^2 = .45, min\ R^2 = .17, \max R^2 = .71, SD = .13$). Figures 25 and 26 show the actual workload reported by pilots p04 and p05 in the Manual, Cueing ON condition to cover the entire scale (1 – 5), as well as the SCOPE estimate.

For pilot p04, it is seen to generally agree with the pilot ratings using the control-related variables, yielding a significant correlation coefficient with $R^2 = .37$, $p < .0001$. The correlation coefficient between reported WL may seem low, but it is important to note that because SCOPE uses control-related inputs, it can't take into account what is happening during hovering stabilization and hover, i.e., the Approach 3 and Hover phases. For pilot p05, the correlation was also significant with P05: $R^2 = .53$, $p < .0001$. The differences in reported WL amplitude and frequency for Pilots p04 and p05, where p04 reporting activity is more frequent and more variable than for p05, have for consequence that the correlation coefficient is paradoxically lower for p04 than for p05.

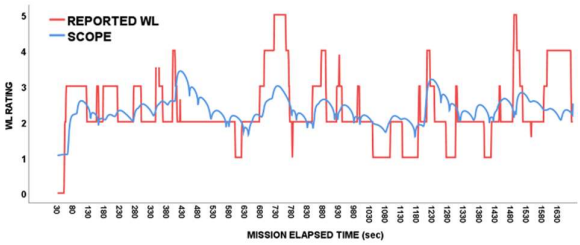


Figure 25. Time series for Observed WL, and SCOPE predicted WL, Pilot p04, Manual, Cueing ON condition.

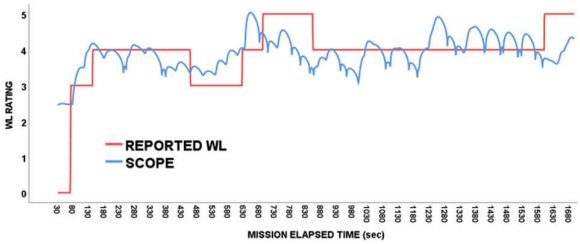


Figure 26. Time series for Observed WL, and SCOPE predicted WL, Pilot p05, Manual, Cueing ON condition.

3.2.2. Dynamic Interface Modeling and Simulation System-Product Metric (DIMSS-PM)

Overall, the DIMSS-PM metric captured the Reported WL relatively well, with a mean $R^2 = .41$, $min\ R^2 = .22$, $\max R^2 = .60$, $SD = .10$.

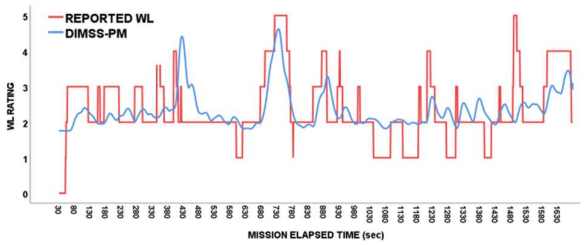


Figure 27. Time series for Observed WL, DIMSS-PM predicted WL, Pilot p04, Manual, Cueing ON condition.

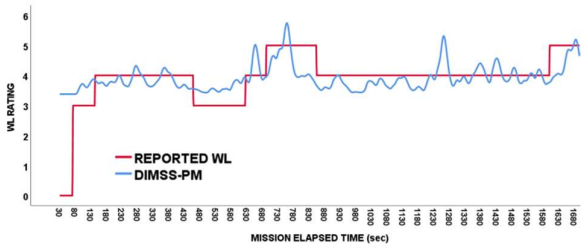


Figure 28. Time series for Observed WL, and DIMSS-PM predicted WL, Pilot p05, Manual, Cueing ON condition.

One can see from the two examples for Pilots p04 (Figure 27) and p05 (Figure 28) (Manual condition, Cueing ON) that the DIMSS-PM metric seems to account better than SCOPE for the effect of Phase, with $R^2 = .37$, $p < .0001$ for pilot p04 and $R^2 = .50$, $p < .0001$ for pilot p05. The same remarks apply to DIMSS-PM regarding the control activity during the phase 3 of the Approach and the Hover phases.

3.2.3. Control Attack Metric

The correlation between Reported WL and Control Attack metric was overall lower than with SCOPE of DIMSS-PM, with a mean $R^2 = .25$ ($min\ R^2 = 0, \max R^2 = 62, SD = .14$). As for SCOPE and DIMSS-PM, the correlation coefficient between the Reported WL and the Control Attack metric was higher for p05 ($R^2 = .31$, $p < .0001$) than for p04 ($R^2 = .11$, $p < .001$), for the same reasons raised earlier.

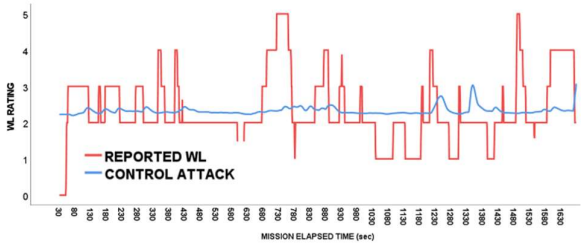


Figure 28. Time series for Observed WL, Control Attack metric predicted WL, Pilot p04, Manual, Cueing ON condition.

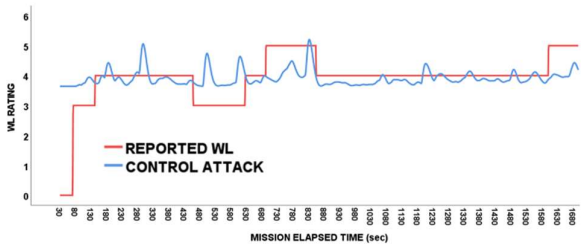


Figure 29. Time series for Observed WL, Control Attack metric predicted WL, Pilot p05, Manual, Cueing ON condition.

3.3. Subjective Metrics

3.3.1. NASA-Task Load Index (TLX)

The NASA TLX (Ref. [36]) is a multi-dimensional rating scale for operators to report their mental workload. It uses six dimensions of workload to provide diagnostic information about the nature and relative contribution of each dimension in influencing overall operator workload. Operators rate the contribution made by each of six dimensions of workload to identify the intensity of the perceived workload.

The operator WL was evaluated along a multidimensional scale with the NASA Task Load Index (TLX) app for iOS using an iPhone. For the pairwise comparison, the participant had to answer 15 pairwise comparison questions to determine weighting of the six subscales: Mental Demand, Temporal Demand, Physical Demand, Effort, Frustration and Performance. Answers from pairwise comparison questions reduce some of the variability associated with the scoring of the individual rating subscales. For the each of the six rating scales, the app displayed instructions on the screen to direct the participant to tap the perceived workload along a scale. This action moves a marker along the scale to mark a workload value along the scale. Available metrics include rating, weight, and adjusted values for each dimension. An overall Weighted Rating is provided. Correlations were computed between the different factors, the overall weighted rating, and the Reported WL.

First, it is interesting to identify the factors contributing the most to the overall Weighted Rating. The Weighted Rating was significantly correlated to Mental Demand, Physical Demand, Performance, and Effort (see Table 8). There was no significant correlation with Temporal Demand and Frustration. Effort was the strongest contributor to the overall Weighted Rating, followed by, in order, Performance, Mental Demand and Physical Demand. Initially, the factorability of the 6 NASA-TLX items was examined. Several well-recognized criteria for the factorability of a correlation were used. Firstly, it was observed that all the items correlated at least .54 with at least one other item, suggesting reasonable factorability.

Table 8. Correlations Matrix for the significant NASA-TLX dimensions and Weighted Rating.

	Weighted Rating	
	R^2	p
Mental Demand	.69	<.0001
Physical Demand	.51	.007
Performance	.76	<.0001
Effort	.96	<.0001

Table 9. Communalities for NASA-TLX dimensions. Extraction Method : Principal Comment Analysis (PCA)

	Initial	Extraction
Mental Demand	1	.601
Physical Demand	1	.542
Temporal Demand	1	.620
Performance	1	.771
Effort	1	.885
Frustration	1	.488

Table 10. Rotation Component Matrix for NASA-TLX dimensions. Extraction Method: Principal Comment Analysis (PCA). Rotation Method: Varimax with Kaiser Normalization.

	Component	
	1	2
Mental Demand	.747	-.209
Physical Demand	.525	.516
Temporal Demand	-.039	.786
Performance	.876	.063
Effort	.927	.158
Frustration	-.003	-.699

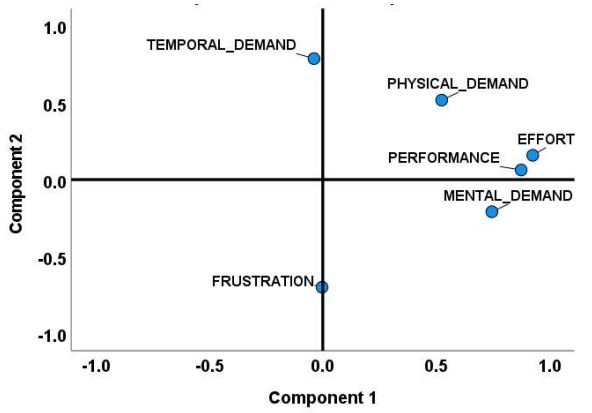


Figure 30: Component Plot in Rotated Space for NASA-TLX dimensions. Extraction Method: PCA. Rotation Method: Varimax with Kaiser Normalization.

Secondly, the Bartlett's test of sphericity was significant ($R^2_{15} = 48.42. p < .001$).

Finally, the communalities were all above .48 (see Table 9), further confirming that each item shared some common variance with other items.

Principal components analysis was used because the primary purpose was to identify and compute composite scores for the factors underlying NASA-TLX. Initial eigen values indicated that the first two factors (components) explained 41%, and 24% of the variance respectively. The solutions for the two factors were each examined using a varimax rotation of the factor loading matrix. Composite scores were created for the two factors, based on the mean of the items which had their primary loadings on each factor. Higher scores indicated greater effect. For Component 1, Effort, and performance had the higher scores. For Component 2, Temporal demand had the highest score.

Overall, these analyses indicated that two distinct factors were underlying pilots' responses to the NASA-TLX and that these factors were overall internally consistent for Component 1 (significance of the correlations between Performance, Effort, and Mental Demand). For component 2, Temporal Demand and Frustration were not significantly correlated. The Reported WL was not significantly correlated to the Weighted Rating ($R^2 = .32, p < .10$). The only significant correlation between Reported WL was with Temporal Demand ($R^2 = .60, p < .001$).

3.3.2. Bedford Rating Scale

The Bedford Scale is a uni-dimensional rating scale designed to identify operator's spare mental capacity while completing a task. The single dimension is assessed using a hierarchical decision tree that guides the operator through a ten-point rating scale, each point of which is accompanied by a descriptor of the associated level of workload. Overall, correlations between Reported WL and Bedford rating was not significant ($R^2 = .08, p = .68$).

DISCUSSION

The aim of this first Operator State Monitoring (OSM) study was to:

1) Select and collect relevant measures of Workload (WL) and Situation Awareness (SA) for the development of real-time OSM.
2) To develop a methodology for the analysis of time series data, in order to identify the local WL triggers, quantify the contribution of each metric to the assessment of WL and SA, evaluate each factor communality between and within measures.
3) To propose foundational elements for the modelling of the relationships between WL, SA, and operative performance.

The validity of a subset of the selected measures of workload was confirmed for the on-line Reported WL and for two of the control metrics, SCOPE, and

DIMSS-PM in the Manual control condition (MAA uncoupled).

The Hypothetical WL Model: External WL

The real-time Reported WL showed to be an accurate representation of the perceived WL under the assumption that the pilots comply with the requested regular response to the prompts. Indeed, for pilots reporting the WL accordingly, the responses were significantly correlated to the Hypothetical WL. This also validates the selection of a 30 sec interval between prompts to provide sufficient temporal resolution for the estimate. Phase relationships between the components remains to be explored. But, one of the areas where a higher sampling rate might prove particularly useful is in the assessment of the accuracy of the component phase values.
This first iteration of the model used three parameters, Phase, Terrain, and Events, combined in an additive manner. The parameter that contributed most to the estimate was Phase, followed by Terrain, then Events. It is important to note that the correlation between Hypothetical WL and Reported WL never exceeded .36, which indicates that other parameters need to be identified and/ or the relationship between the present parameters refined, such as adjusting weights and/ or phases. For example, speed, altitude and route geometry are additional parameters of interest.

However, interindividual variability can't be ignored, and the relative contribution of each parameter is susceptible to vary as a function of pilot expertise and/ or sensitivity to his/her own internal states.

Using a Principal Component Analysis (PCA), two components were identified that contributed to the Hypothetical WL. The first component related to Terrain and Phase, which are negatively correlated. The use of Attitude Command likely contributed to the impact of Terrain and Phase of Flight in the Hypothetical WL estimate. The second component related to Events. This parameter didn't contribute as expected to the total Hypothetical WL. This may be the result of multiple factors, such as pilot deviation from the commanded altitude (flying higher), pop-up obstacle placement offset relative to the flight path, or the simulation environment/ conditions missing additional task loading or compelling realism.

Control Metrics: Internal WL

SCOPE and DIMSS-PM correlations with Reported WL were overall relatively similar and relatively high, with respectively $R^2 = .45$ and $R^2 = .41$ on average. Meanwhile, the correlation between Reported WL and Control Attack metric was much lower ($R^2 = .24$). Qualitatively speaking, SCOPE and DIMSS-PM seemed to capture different aspects of the control activity. For the two reported cases (Pilots p04 and p05), DIMSS better captured the higher values of Reported WL.

Subjective Metrics

For the subjective metrics, NASA-TLX Weighted Rating showed no significant correlation with the Reported WL. The only observed significant correlation was with Temporal Demand. Meanwhile the Bedford ratings didn't correlate with the mean Reported WL values. First these measures are off-line, a posteriori measures of WL, and therefore, susceptible to noise and averaging out the information, or reflecting a recency effect.

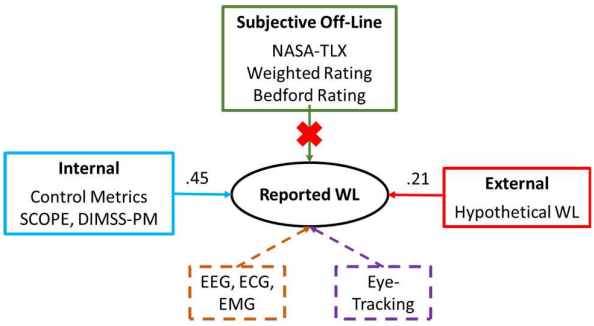


Figure 31. Summary of the correlations between observed variables.

CONCLUSIONS

The partial, preliminary results suggest that control input and real-time WL report are valid measures of WL for this specific combination of task and environment in simulation and for the conditions and the pilot data analyzed thus far.

The next steps will include:

- The analysis of the data for the Autonomous Flight (MAA coupled) condition
- The analysis of performance data [reaction time to events (obstacles and ASE events, CFIT, WL prompts), technical performance measures (lateral, vertical, and heading deviation from flight path, speed deviation, ability to recover intended position/speed control, radial error for Hover]
- EEG, ECG, and eye-tracking data analyses. Comparisons with the previously processed data will be made by the reduction of the data through time windowing ($\cong 30\ sec$) indexed by the time of the WL prompts and plunges
- Comparison of WL and SA between Manual and Autonomous flight conditions, Cueing conditions
- Correlations between variables within and between categories (Performance metrics and Biometrics)
- Integrate the audio and tactile cueing into SCOPE.

A second experiment is planned that includes:

- The same routes, events, mission (MEDEVAC), and visibility level
- A different pool of pilots (20+ Army pilots 10th Combat Brigade)
- Secondary Task: Visual and/or auditory instruction to change radio frequency
- Tertiary task: Target detection and identification of vehicle on ground
- Same biometrics + EMG
- HMD: SA Photonics + demonstration with Elbit X-sight
- A different eye-tracking system (Ergoneers Dikablis) mounted on the HMDs

If successful in demonstrating the validity of the WL measures over all conditions and pilots, a modelling effort will be undertaken to integrate real-time WL measurement into the system architecture, including the development of a prototype decision aiding component and potentially scalable displays.

REFERENCES

[1] Edwards, T. (2017, February). The Human Performance Envelope: Past Research, Present Activities and Future Directions. In *FAA Human Performance meeting* (No. ARC-E-DAA-TN39518).

[2] Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2), 159-177.

[3] Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human factors*, *61*(3), 374-392.

[4] Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., & Linton, P. M. (1989). Operator workload: Comprehensive review and evaluation of operator workload methodologies.

[5] Wickens, C. D., Boles, D., Tsang, P., & Carswell, M. (1984). *The limits of multiple resource theory in display formatting: Effects of task integration*. ILLINOIS UNIV AT URBANA-CHAMPAIGN.

[6] Gopher, D., Donchin, E., Boff, K. R., Kaufman, L., & Thomas, J. P. (1986). Handbook of perception and human performance.

[7] Endsley, M. R., & Garland, D. J. (Eds.). (2000). *Situation awareness analysis and measurement*. CRC Press.

[8] Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human factors*, *37*(1), 65-84.

[9] Chauvin, C., Lardjane, S., Morel, G., Clostermann, J. P., & Langard, B. (2013). Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS. *Accident Analysis & Prevention*, *59*, 26-37.

[10] Durso, F. T., & Gronlund, S. D. (1999). Situation awareness. *Handbook of applied cognition*, 283-314.

[11] Hendy, K. C. (1995). *Situation awareness and workload: Birds of a feather?* Defence and Civil Institute of Environmental Medicine.

[12] Wagner, A. R. (1976). Priming in STM: An information processing mechanism for self-generated or retrieval generated depression in performance. In T. J. Tighe & R. N. Leaton (Eds.), *Habituation: Perspectives from child development, animal behavior, and neurophysiology.* Hillsdale, N.J.: Erlbaum, 1976.

[13] Endsley, M. R. (1993). A survey of situation awareness requirements in air-to-air combat fighters. *The International Journal of Aviation Psychology*, *3*(2), 157-168.

[14] Nählinder, S., Berggren, P., & Svensson, E. (2004, September). Reoccurring LISREL patterns describing mental workload, situation awareness and performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 21, pp. 2485-2488). Sage CA: Los Angeles, CA: SAGE Publications.

[15] Doyle, M. J. (2008). Modeling the interaction between workload and situation awareness: an overview and future course. In *Symposium on Human Interaction with Complex Systems and 2nd Sensemaking of Complex Information Annual Conference*.

[16] Hart, S. G., & Wickens, C. D. (1990). Workload assessment and prediction. In *Manprint* (pp. 257-296). Springer, Dordrecht.

[17] De Waard, D., & Brookhuis, K. A. (1996). The measurement of drivers' mental workload.

[18] Anon, “Aeronautical Design Standard, Handling Qualities Requirements for Military Rotorcraft,” US Army Aviation and Missile Command, USAAM-COM, ADS-33E-PRF, 21 March 2000.

[19] Szoboszlay, Z., Miller, J.D., Godfroy-Cooper, M., et al. (2021). The Design of Pilot Cueing for the Degraded Visual Environment Mitigation (DVE-M) System for Rotorcraft. Proceedings of the 77th Vertical Flight Society Annual Forum and Technology Display, Virtual, May 10th-14th. Proceedings of the 77th Vertical Flight Society Annual Forum and Technology Display, Virtual, May 10th-14th.

[20] Miller, J. D., Godfroy-Cooper, M. and Szoboszlay, Z. P., “Augmented-Reality Multimodal Cueing for Obstacle Avoidance: Towards a New Topology for Threat-Level Presentation,” Proceedings of the 75th American Helicopter Society Annual Forum, Philadelphia, PA, May 13-16, 2019

[21] Takahashi, M. D., Fujizawa, B. T., Lusardi, J. A., Whalley, M. S., Goerzen, C. L., Schulein, G. J., ... & Waldman, D. W. (2021). Autonomous Guidance and Flight Control on a Partial-Authority Black Hawk Helicopter. *Journal of Aerospace Information Systems*, *18*(10), 686-701.

[22] Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., ... & Olmstead, R. (2004). Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction*, *17*(2), 151-170.

[23] Freeman, F. G., Mikulka, P. J., Scerbo, M. W., Prinzel, L. J., & Clouatre, K. (2000). Evaluation of a psychophysiologically controlled adaptive automation system, using performance on a tracking task. *Applied Psychophysiology and Biofeedback*, *25*(2), 103-115.

[24] Watford, T. S., O’Brien, W. H., Koerten, H. R., Bogusch, L. M., Moeller, M. T., Sonia Singh, R., & Sims, T. E. (2020). The mindful attention and awareness scale is associated with lower levels of high-frequency heart rate variability in a laboratory context. *Psychophysiology*, *57*(3), e13506.

[25] https://humansystems.arc.nasa.gov/groups/TLX/tlxapp.php

[26] Padfield, G. D., et al. "Where does the workload go when pilots attack manoeuvres? An analysis of results from Flying qualities theory and experiment." European Rotorcraft Forum, 20 th, Amsterdam, Netherlands. 1994.

[27] Memon, W. A., et al. "Helicopter handling qualities: A study in pilot control compensation." The Aeronautical Journal 126.1295 (2022): 152-186.

[28] Roscoe, M. F. and Wilkinson, C. H., DIMSS – JSHIP'S Modeling and simulation process for ship/helicopter testing and training, AIAA Modeling and Simulation Technologies Conference and Exhibit, Monterey, 2002.

[29] Bachelder, Edward, and Bimal Aponso. "A Theoretical Framework Unifying Handling Qualities, Workload, Stability, and Control." Vertical Flight Society 77th Annual Forum & Technology Display, Virtual. 2021.

[30] Norwich, K. H. (1987). On the theory of Weber fractions. *Perception & Psychophysics*, *42*(3), 286-298.

[31] Bachelder, Edward N. "SCOPE-Pilot Workload Estimation Using Control Response: Theoretical Development and Practical Demonstration." AIAA Scitech 2020 Forum. 2020.

[32] Bachelder, Edward N., and Martine Godfroy-Cooper. "Pilot Workload Estimation: Synthesis of Spectral Requirements Analysis and Weber’s Law." AIAA Scitech 2019 Forum. 2019.

[33] Bachelder, E., Berger, T., Godfroy-Cooper, M. Aponso, B. "Pilot Workload and Performance Assessment for a Coaxial-Compound Helicopter and Tiltrotor During Aggressive Approach." Vertical Flight Society 77th Annual Forum & Technology Display, Virtual. 2021.

[34] Bachelder, Edward N., Lusardi, J., Aponso, B., Godfroy-Cooper, M. "Estimating Handling Qualities Ratings from Hover Flight Data Using SCOPE." AIAA Scitech 2021 Forum. 2021.

[35] Hart, Yuval, et al. "Comparing apples and oranges: fold-change detection of multiple simultaneous inputs." PloS one 8.3 (2013):e57455.

[36] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.