Perceptual-Cognitive & Physiological Assessment of Training Effectiveness

Conference Paper · November 2017

Thomas Romeas
L'Institut national du sport du Québec (INS Québec)

Jocelyn Faubert
Université de Montréal

Perceptual-Cognitive & Physiological Assessment of Training Effectiveness

Jaclyn Hoke[1], Christopher Reuter[2], Thomas Romeas[3,4], Maxime Montariol[2], Thomas Schnell[2],
Jocelyn Faubert[4,5]

[1]Rockwell Collins, Inc., Cedar Rapids, IA
[2]University of Iowa Operator Performance Lab, Iowa City, IA
[3]École de Technologie Supérieure, Montreal, Qc, CA
[4] CogniSens Applied Research Center, Montreal, Qc, CA
[5] University of Montreal, School of Optometry, Faubert Lab, Montreal, Qc, CA

jaclyn.hoke@rockwellcollins.com, christopher-m-reuter@uiowa.edu, tromeas@c-arc.org,
maxime-montariol@uiowa.edu, thomas-schnell@uiowa.edu, jocelyn.faubert@umontreal.ca

ABSTRACT

Several trends within the simulation and training industry are emphasizing the need for measurable proof that training solutions meet or exceed the requirements for delivering effective training. Cognitive state is a key component of learning, meaning that classification of cognitive state and capacity can provide a measure of training effectiveness. However, accurate classification of trainee state is an extremely challenging task. The more traditional subjective assessment methods have several limitations, while objective assessment methods can be difficult to implement.

We conducted an exploratory study that evaluated the cognitive and physiological load engaged during flight simulation and live flight during maneuvers of three levels of difficulty. The study represents the work performed to date in the first year of a multi-year effort to design a method for assessing the efficacy of training content and devices, including live platforms, that is based on objective cognitive state assessment techniques coupled with control input and mission/platform performance measures. The method employs NeuroTracker, a validated tool for evaluating or training perceptual-cognitive skills, to measure spare cognitive capacity, and physiological measures of workload based on analysis of eye tracking and electrocardiogram data.

This paper briefly summarizes the design, implementation, and initial results of this study. It summarizes the next steps required to further refine the proposed method for assessing training efficacy and describes the planned follow-on effort. Finally, it discusses additional applications of this method in military and commercial training markets, such as the real-time adaptation of training content to trainee skill level and state.

ABOUT THE AUTHORS

Dr. Jaclyn Hoke is a Principal Engineer in the Rockwell Collins Advanced Technology Center with 10 years of experience in research and development. Jaclyn has an undergraduate degree in Applied Mathematics and an MSc and PhD in Computer Engineering. She is the technology lead for Rockwell Collins research initiatives in Cognitive State Assessment and Training Effectiveness, and served as the I/ITSEC Operation Blended Warrior Special Event IPT Lead for Performance Measurement. Prior research focuses include developing architectures and maturing technologies for Live Virtual Constructive distributed training, and the design and human factors assessment of future avionics technologies and pilot aids.

Mr. Christopher Reuter is a researcher at the University of Iowa Operator Performance Laboratory (OPL). Chris has a Bachelor of Science degree in Mathematics, and is currently pursuing his MSc degree in Industrial Engineering with an emphasis in human factors in aviation from the University of Iowa. He joined the OPL team in 2012. Chris's role at OPL includes executing and directing human factors studies, as well as performing data collection and data analysis.

Dr. Thomas Romeas is an industrial post-doctoral researcher at the École de Technologie Supérieure (Department of Software and IT Engineering) and a scientific project manager at the CogniSens Applied Research Center. His interests lie mainly in applied science and in the specialized field of vision, cognition, brain injury or human performance. He is currently developing cognitive training programs for athletes and a number of other populations using virtual reality and motion capture, as well as other cutting edge technologies. He received his PhD in vision neurosciences and his MSc in neuropharmacology from the University of Montreal, in addition to a BSc in physiology (University Lyon 1, FR).

Mr. Maxime Montariol is a student in aerodynamics with a specialization in Human Factors at the École de l'Air. He holds an undergraduate degree in Aerodynamics, and is currently studying at the OPL to further his knowledge of Human Factors related to aeronautics. After completing his study at the OPL, he will hold a Master's degree in Aeronautics and Aerospace. His goal is to specialize in Human Factors in military aeronautics.

Dr. Tom "Mach" Schnell is a Professor in Industrial and Mechanical Engineering with a specialization in Human Factors/Ergonomics at the University of Iowa. He has an undergraduate degree in EE and an MSc and Ph.D. in Industrial Engineering with a specialization in Human Factors. He has lead over 250 human factors research projects. He is a Commercial pilot, research test pilot, and flight instructor with helicopter, jet, and glider ratings and has been actively involved in flight testing using fixed wing and helicopter testbeds as a pilot and principal investigator since 2004.

Dr. Jocelyn Faubert is a full professor and directeur adjoint (equivalent of vice dean) of research and graduate studies at the School of Optometry of Université de Montréal. He holds the NSERC-Essilor Industrial Research Chair on Visual Function (since 2003). He is also a founder and the chief science officer of CogniSens and a not for profit organization 'The CogniSens Applied Research Center' that is focused on the translational research related to the use of the NeuroTracker System and other technologies originally developed in his laboratory. His research interests are diverse dealing from low- to high-level perceptual and cognitive processing, human performance, development, aging, optics and photonics spanning from a fundamental perspective to translational research. He has 19 distinct patents & patents pending spanning from biophotonics, brain imaging, optical designs, and behavioral procedural methods for which the great majority have been licensed or sold to companies such as Essilor, CogniSens and Annidis health.

Perceptual-Cognitive & Physiological Assessment of Training Effectiveness

Jaclyn Hoke[1], Christopher Reuter[2], Thomas Romeas[3,4], Maxime Montariol[2], Thomas Schnell[2], Jocelyn Faubert[4,5]

[1]Rockwell Collins, Inc., Cedar Rapids, IA
[2]University of Iowa Operator Performance Lab, Iowa City, IA
[3]École de Technologie Supérieure, Montreal, Qc, CA
[4] CogniSens Applied Research Center, Montreal, Qc, CA
[5]University of Montreal, School of Optometry, Faubert Lab, Montreal, Qc, CA

jaclyn.hoke@rockwellcollins.com, christopher-m-reuter@uiowa.edu, tromeas@c-arc.org, maxime-montariol@uiowa.edu, thomas-schnell@uiowa.edu, jocelyn.faubert@umontreal.ca

INTRODUCTION

Continued budgetary pressures and rapid advancements in modeling and simulation capabilities are fostering substantial growth in the use of simulation-based training technologies in the Department of Defense. Additionally, low-cost, commercial-off-the shelf (COTS) technologies are beginning to supplant custom simulation environments. The government and industry continue to invest in these technologies, driven by the claims that these capabilities will save money without compromising training efficacy. Yet the Government Accounting Office has found that appropriate measures to support these assertions do not exist from either a cost or effectiveness standpoint (U.S. General Accounting Office, 2013). The need for measurable proof that training solutions, custom or COTS, meet or exceed the requirements to deliver effective training is taking on greater emphasis. Current measures of training success are typically a matter of subjective assessment by an instructor, and readiness is often determined in terms of time (e.g., flight hours) and qualification tasks. However, these measurement approaches highlight task outcome, as opposed to assessing knowledge and skill acquisition by the trainee.

Cognitive state is a key component of learning (e.g. Cognitive Load Theory – Sweller, 1988), meaning that classifying the cognitive demand of a simulation may be one of the most important criteria to evaluate the effectiveness of a training device. Yet the accurate classification of a trainee’s cognitive state is an extremely challenging task. It is traditionally performed using subjective rating scales, such as the NASA Task Load Index (Hart and Staveland, 1988), but these methods can be disruptive to the primary training task. A subjective assessment can also be highly variable in its outcome due to biases created by vested interests (Schreiber et al., 2006). An alternative approach is to employ objective assessment methods that collect data from sensors on, or about, the trainee and their environment without interrupting the training task. Objective measures fall into three categories: control input based measures, performance based measures, and physiological measures. These measures may be more desirable but are often more difficult to implement. In this study, we performed an initial investigation into objective trainee state assessment using perceptual-cognitive techniques and physiological measures. The long term goal of the research is to develop a method for assessing the efficacy of training content and devices, including live platforms.

One model that provides a foundation for our study design is Cognitive Load Theory, which states that there are three sources of cognitive load competing for resources: intrinsic load, germane load, and extraneous load (Sweller, 1988). Intrinsic load is the cognitive load associated with the complexity of the task being trained, and characterizes the resources required to understand the training materials. Germane load is the cognitive load associated with comparing new training tasks to prior knowledge and translating it into schemas in long-term memory. Transferring information into long-term memory is linked to learning. Finally, extraneous load is the cognitive load associated with processing information not related to learning and detracts from skill acquisition during training. In terms of the cognitive assessment techniques as employed in this study, physiological based measures were used to quantify the total load while perceptual-cognitive measurement techniques served as a secondary task to provide insight into the spare capacity that can be dedicated to the primary task. This configuration was chosen to assess task sharing in the cockpit with flying (aviate) as the primary task and the perceptual-cognitive task being a configurable, consistent substitute for secondary cockpit tasks (navigate, communicate).

Perceptual-Cognitive Measurement

Perceptual-cognitive skills is a term that refers to the role played by both perceptual and cognitive processes required to extract meaningful contextual information from a dynamic scene and support decision making. A few years ago a new perceptual-cognitive methodology was introduced in the field of sport performance (Faubert and Sidebottom, 2012). The method, called Three Dimensional Multiple Object Tracking (3D-MOT), is commercially available under the name of NeuroTracker (NT). The technique is a highly levelled perceptual-cognitive task which stimulates a high number of brain networks that must work together during the exercise including complex motion integration, dynamic, sustained and distributed attention processing and working memory. The task targets fundamental cognitive functions that are required during sport practice but also in everyday life. Consequently, the paradigm has demonstrated evidence of transfer across different populations and domains. The main advantages of this non-contextual technique are that it is simple to use, allows major gains with minimal training time and can be generalized to a variety of environments.

The 3D-MOT training technique has revealed superior skills in elite athletes compared to sub-elites and novices for learning this complex and neutral dynamic visual stimulus (Faubert, 2013). Another study revealed that 3D-MOT performance was most likely related to the athletes' ability to see and respond to various stimuli on the basketball court (Mangine et al., 2014). Outside of the sport-performance domain, the technique was also characterized as the best measure, among other predictors, to predict time completion and accuracy of a simulated laparoscopic surgery task in medical surgeon students (Harenberg et al., 2016). Furthermore, recent neurological evidence has demonstrated that 10 sessions of 3D-MOT training improved attention, visual information processing speed and working memory recorded through neuropsychological tests and quantitative electroencephalography in healthy young adults (Parsons et al., 2014). These findings were supported by another study showing significant improvement on neuropsychological working memory tests in soldiers following 3D-MOT training compared to an active control group and a passive control group (Vartanian et al., 2016). Additionally, the methodology has shown transfer benefits to socially relevant tasks such as biological motion perception in the elderly (Legault and Faubert, 2012) and to real-game situations (passing decision-making) in soccer players (Romeas et al., 2016).

This evidence suggests that the 3D-MOT measure of performance is an appropriate marker of human perceptual-cognitive skills. Moreover, the simplicity of the technique, which only requires a screen and a computer to be executed, and the benefit of objective thresholds obtained from the method, makes it adaptable to multiple contexts including very complex environments. For these reasons, this perceptual-cognitive paradigm is employed in the context of this study.

Physiological Assessment

Understanding and monitoring the changes in the cognitive workload in pilots can offer critical quantitative information about their progression and performance while undergoing training. Unfortunately, accurate real-time objective quantification of cognitive workload using unobtrusive physiological signals has not been achieved. Unobtrusive in this context refers to both the ease of the physiological sensor montage and non-interference of the workload assessment system with the primary task. The Cognitive Assessment Tool Set (CATS) system is worn under the pilot's flight suit and does not interfere with the flying task, yet it provides a real-time assessment of cognitive workload to allow performance assessment and training adaptations.

In well over a decade of physiological based assessment work, we investigated many sensors and arrived at the conclusion that the electrocardiogram (ECG) waveform is by far the best signal for workload assessment. ECG signals respond strongly to changes in the limbic system affected by a person's attention and cognitive engagement. We use a deterministically nonlinear dynamical classifier to assess cognitive workload with great success (Engler et al., 2013; Schnell and Engler, 2014). The research community has known for a number of years that human physiological signals in general, and ECG specifically, are deterministically nonlinear (also known as chaotic) systems (Govindan et al., 1998; Kozma, 2002; Owis et al., 2002). Chaotic systems are often not well represented via the normal scalar time series. The dynamics of the system are obscured in the single dimension, but become apparent when transformed into a multi-dimensional embedded phase space (Richter and Schreiber, 1998). We refer to this approach as the Chaotic Physiological Classifier (CPC) method. Figure 1 depicts an example of this technique, where a portion of an ECG signal recorded during a recent study is embedded in phase space and is then coarse-grained into a numerical array that represents a quantitative signature of operator state.

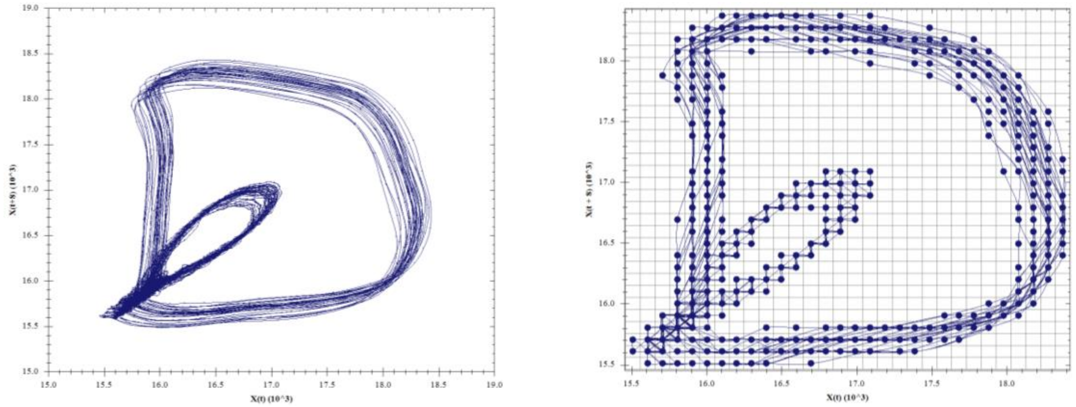

Figure 1. Example Scalar ECG Transformed into Embedded Phase Space (Left) and Coarse-Grained (Right)

Study Goals

The overarching goal of this multi-year research effort is to establish and validate physiological measures of cognitive workload that characterize skill acquisition. It is our fundamental hypothesis that this relationship exists, and that a combination of physiological measures and technical flight performance can be used to assess the efficacy of training devices and content in the transfer of requisite skills. This paper reports on the work performed during the first year of the effort. It describes the initial study that has established our experimental approach where physiological assessment techniques and 3D-MOT methods are combined within an experimental aircraft testbed to assess the performance of pilots in simulated and live flight.

METHOD

Participants

10 low-time (100-300 total flight hours) evaluation pilots (EP) volunteered to participate in the study. The low sample size was used initially to develop a proof-of-concept and justify follow-on research with a larger, more diverse sample size. All subjects were males between the ages of 20 and 25 holding a valid U.S. private pilot certificate and, at a minimum, a current Class III medical. The pilots were asked to have regular access to a personal computer to support the NT home training.

Testbed

The study was conducted in the University of Iowa Operator Performance Laboratory's (OPL) experimentally rated Aero Vodochody L-29 jet trainer, shown in Figure 2. It is a fully acrobatic platform capable of performing high dynamic maneuvers up to +8/-4gz at speeds up to Mach 0.7. This single engine, tandem-seat aircraft is equipped with human performance state assessment tools to monitor the EP physiological based cognitive workload parameters, control inputs, and six channel audio/visual recording capabilities for human factors assessments. The aircraft is instrumented in such a manner that it also can serve as an "aircraft-in-the-loop" (AIL) simulator. This capability was extensively used to familiarize the EPs with the display symbology and crew duties. The AIL configuration was used to perform the flight maneuvers during the simulator portion of the test run.

Figure 2. Aero Vodochody L-29 Jet Trainer

The EP performed the flight maneuvers from the aft cockpit of the L-29, shown in Figure 3, during both the AIL simulation and live flight ("jet") experimental conditions. The NT software was integrated into the testbed, and

Figure 3. Aft Cockpit with NeuroTracker

displayed on the upper screen. The EP interacted with the NT system via the display's touch interface. The primary flight display (PFD) used for flight instrument tasks and a timer were depicted on the lower display surface. The timer was required to ensure the baselining task and certain flight maneuvers were performed for the specified duration. The CATS human performance assessment suite collected flight technical and physiological data from various sensors and simulator systems within the testbed. The tool also provided a user interface for the Experiment Payload Controller (EPC) to start and stop the test run or tag noteworthy events. Data from the sensors was timestamped, synchronized, and recorded to a database to support data analysis. CATS features tabs that were used to view the data while it was being recorded, including the "Vitals" screen display real-time ECG data and the "Video" screen displaying eye gaze information.

Measures

Several different measures, including both objective and subjective data, were collected during the study. Objective measures include physiological and flight technical measures collected by the CATS system and the perceptual-cognitive measures produced by NT. We also collected subjective cognitive workload data and situation awareness (SA) ratings. The specific measures in each of these categories are described in detail here.

Perceptual-Cognitive Measures. A standalone and a cloud version of the NT system were used during the experiment. In both cases, the 2D 'CORE' mode of the NT was used (Figure 4). The exercise required tracking four targets among eight spheres projected within a cube space, subtending a visual angle of 30° in the context of this study. The spheres followed a linear trajectory in the space. Deviation occurred only when the balls collided against each other or the walls. In order to support an effective distribution of attention, a fixation spot was presented in the center of the cube throughout the experimentation. A typical session, based on a staircase procedure, lasted about 6 minutes and included 20 trials. The staircase procedure consists of increasing speed displacement if the subject correctly identified all of the indexed targets or decreasing speed displacement if at least one of the targets was missed. Displacement thresholds were then evaluated using a 1-up 1-down staircase procedure (Levitt, 1971). After each correct response, the dependent variable (speed of ball displacement) would increase by 0.05 log and decrease by the same proportion after each incorrect response, resulting in a threshold criterion of 50%. The threshold was estimated by the mean of the speeds of the last four inversions in cm/s.

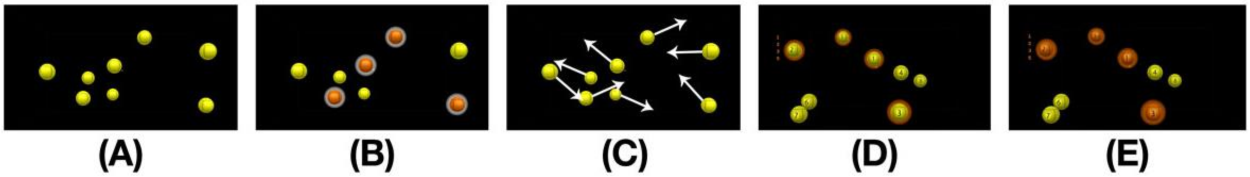

Figure 4. NeuroTracker 'CORE' mode: (A) Target Presentation, (B) Target Identification, (C) Displacement, (D) User Response, and (E) Feedback

Physiological Measures. Two specific physiological measures were collected for processing using the CATS system during the study: ECG signals and eye gaze. Previous research analyzing the entire ECG waveform using the CPC method has revealed a very high correlation with cognitive workload. ECG data was collected using three-lead electrodes and the NeXus-4, a portable, wireless Bluetooth biofeedback device. Eye gaze analysis is used to detect inattention or cognitive tunneling. In this study we employed a simple measure of the percentage of time the participants spent viewing designated areas of the cockpit, specifically the upper display showing NT versus the lower display showing flight instruments (summarized in Table 1). This data was collected using a Dikablis eye tracker installed on the visor of the helmet. An additional scene camera was placed at the top of the helmet to provide a forward looking view.

Table 1. Eye Gaze Measures

Upper Display Intersection (%)	Defined as the percentage of time that a participant's eye gaze intersected with the upper display. The upper display was only used for the NT task.
Lower Display Intersection (%)	Defined as the percentage of time that a participant's eye gaze intersected with the lower display. The lower display contained the PFD and timer.

Flight Technical Measures. Performing the experimental flight maneuvers was the EPs primary task during this study. Collecting flight technical measures assessed the participant's ability to remain within the ideal threshold when performing a maneuver. These measures also indicated if the EP shed the primary task (flying) to perform the NT secondary task. The flight technical measures recorded during the study are summarized in Table 2.

Table 2. Flight Technical Measures

Altitude Error (ft)	Altitude error was defined as the deviation of the target altitude, within the tolerance defined by the maneuver. Root mean square (RMS) was calculated for all altitude errors in the duration of the flight maneuver.
Roll Error (deg)	Roll error was defined as the deviation of the bank (roll) angle from the selected bank angle, within the tolerance defined by the maneuver. RMS was calculated for all roll error samples in the duration of the flight maneuver.
Vertical Speed Error (ft/min)	Vertical speed error was defined as the deviation in the rate of climb or descent (500 ft/min) for the High maneuver. RMS was calculated for all vertical speed error samples in the duration of the High maneuver.

Subjective Evaluation. The subjective assessment of cognitive workload was performed using a 10-point Bedford Workload Scale (Roscoe, 1984). The EP's subjective SA was assessed using the 3D Situation Awareness Rating Technique (SART). SART (Taylor, 1989) uses three scales, each rated from 1 (low) to 7 (high), which are combined to form a single SA metric. The three scales are *Demand on Attentional Resources* – how much attention and effort was required to perform the primary task, *Supply of Attentional Resources* – how much spare attention was available for secondary tasks, and *Understanding of the Situation* – the overall understanding of what happened during the scenario. The SART score is then calculated as SART = Understanding – (Demand – Supply).

Procedure

Orientation. Orientation visits were conducted in groups of three or four participants. The purpose of this visit was to review and sign the Institutional Review Board Informed Consent Document forms and receive an introductory briefing. The introductory briefing presented information regarding the purpose of the study, the registration process for the home-based NT sessions, the duration of these training sessions, and the flight technical and physiological measures being collected. Following the briefing, the EPC provided instructions on a laminated chart to each EP on how to attach the ECG leads to their body. After the EPs successfully placed the leads, participants completed an initial baseline session of NT on their own laptop. This step was performed to ensure they could use the NT software at home. After each NT session, the participants self-reported Bedford workload ratings.

NT Gym Consolidation. The home-based NT training, called "gym consolidation", was performed at home. Participants were required to perform 15 NT sessions, each consisting of 20 trials that lasted 8 seconds in duration. Every two to three days, participants completed 3 consecutive NT sessions on their own computer within a three-week span. The procedure of consolidation familiarized the participant with the task in order to attain stable thresholds of performance before combining the exercise with a flight situation. Establishing this baseline metric allows the identification of the cognitive loading associated with the flight tasks.

Simulator and Flight Test. The second phase was completed on an individual basis. The EP arrived at OPL at approximately 8 A.M. on the scheduled day of the flight and was escorted to the briefing room. The EPC briefed the subject on the objectives for the day, the operation of the NT software during flight, and provided an in-depth review of the flight maneuvers. This initial briefing was followed by a detailed safety briefing and egress training by the safety pilot (SP), attachment of the ECG leads, and calibration of the eye tracking system.

Once the ECG was successfully attached and the eye tracking system was calibrated, the flight maneuvers were flown in the AIL simulator configuration, followed by in actual flight in the L-29 jet trainer. Each EP performed three levels of flight maneuvers to assess their flight technical performance while using the NT software as a secondary task to provide a measure of spare cognitive resources. Each participant also flew the three flight maneuvers without the NT present. The maneuvers, shown in Figure 5, consisted of climbs, descents, and turns that are similar to those flown during normal pilot training. The order of these scenarios was randomized for both the simulator and flight portions of the study. In all cases, simulator maneuvers were completed prior to flight for all EPs due to safety considerations with low-time pilots that lack experience in the experimental platform.

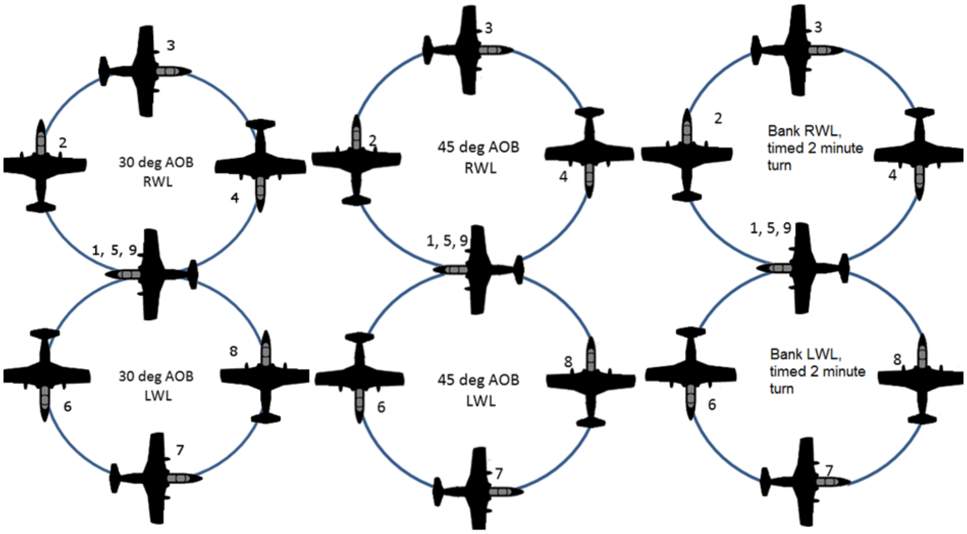

Figure 5. (Left to right) Low, Medium, and High Difficulty Flight Maneuvers

The Low and Medium difficulty maneuvers required the EP to perform a 360° right turn while maintaining an altitude target within a specified threshold. The target thresholds were ±200 feet and ± 100 feet for the Low difficulty and Medium difficulty maneuvers, respectively. Once the heading assignment was met, ±10°, the EP began a 360° left turn. In addition to maintaining altitude and meeting the heading target, the EP was also required to maintain a specified angle of bank (AOB) throughout the turn. The target AOB was 30° for the Low difficulty maneuver and 45° for the Medium difficulty maneuver. The EP continued making these turns until the NT trials were complete or once the cycle of the maneuver was complete when the NT task was absent.

The High difficulty maneuver began with a 1000 ft climb at a rate of 500 ft/min while simultaneously rolling right to accomplish the complete 360° turn in 2 minutes (stable, timed turn) to return to the original heading, followed by a descent -500 ft/min to the initial altitude while performing a timed 2 minute 360° turn to the left. To increase difficulty, the turn rate indicator was absent from the EP's cockpit displays. A timer was added to the lower corner of the PFD, and their goal was to meet both the heading and altitude targets at exactly 120 seconds (theoretically creating a perfect 3dps roll rate). Once the heading target was met, the EP leveled out, reset the timer, and began a 3 dps left turn descending at 500 ft/min to reach a target altitude of 1000 ft below the new starting altitude and target heading of the original heading in exactly 120 seconds. The EP continued this maneuver until the NT trials were complete or once the cycle of the maneuver was complete when the NT condition was absent.

After the EP accepted control of the aircraft, using pre-briefed procedures, the EP was given a small amount of time to familiarize himself with flying the platform (both in AIL simulation and live flight). Following the familiarization run, the EP flew a baseline task for approximately one minute without NT present to establish an initial physiological baseline. Additionally, because the difficulty of the previous tasks, unexpected maneuvers, or long waits due to traffic (in live flight) may induce emotional responses, the EP performed a thirty second "Eyes Closed" task (with the SP in control of the aircraft) prior to each primary task in the run matrix. This allowed the baseline physiological metrics to adapt organically to the environment. The EP's specific run order, including the initial baseline, each maneuver, and the "Eyes Closed" task, was provided on a kneeboard card. The kneeboard also provided the sequence of steps for each maneuver in case the EP did not recall the steps from the initial briefing.

RESULTS

NeuroTracker. In order to evaluate the impact of the aviation task on NT performance, a paired Student-t test was performed on NT displacement thresholds during the latest home training session (mean score of sessions 13 – 15) and the flight condition (mean score of sessions done while flying the simulator and jet). Moreover, a within-subjects repeated-measures two-way analysis of variance (ANOVA) was applied to the NT displacement thresholds. This analysis compared the main effects of the test apparatus (simulator, jet), maneuver's level of difficulty (Low, Medium, High) and interaction between both factors. Pairwise comparisons were made using Bonferroni corrections in order to observe differences between the maneuver's level of difficulty. For each analysis, a Mauchly's test and a Shapiro-Wilk test were performed to control for sphericity ($p > 0.05$) and normality ($p > 0.05$) respectively.

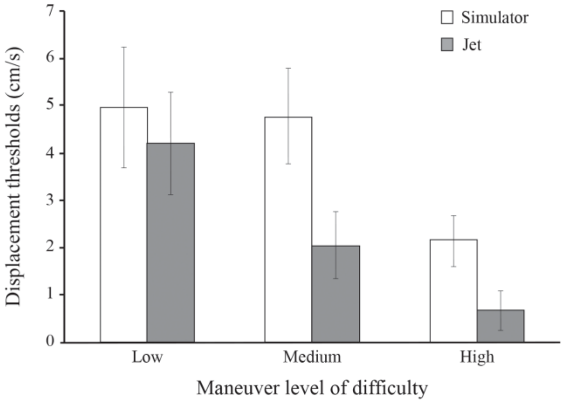

Figure 6. NeuroTracker Performance in the Simulator and Jet Across Maneuvers

The paired Student-t test revealed a strong significant difference between the NT displacement thresholds when performed alone versus in combination with flight maneuvers (t[9] = 12.626, $p < 0.001$). The ANOVA results demonstrated a significant main effect of the test apparatus ($F_{1,9} = 10.202$, $p = 0.011$, $\eta^2 = 0.531$), indicating lower overall performance on NT in the jet compared to the flight simulator (Figure 6). There was also a significant main effect of the maneuver's level of difficulty on NT displacement thresholds ($F_{2,18} = 7.279$, $p = 0.005$, $\eta^2 = 0.447$). Pairwise comparisons revealed a significant difference in the thresholds between the 'Low' and 'High' level of difficulty ($p = 0.039$), an almost significant difference between the 'Medium' and 'High' level of difficulty ($p = 0.073$) and no significant difference between the 'Low' and 'Medium' level of difficulty ($p = 0.4$). There was a trend for an interaction between test apparatus and the maneuver's level of difficulty which failed to reached the expected level of significance ($F_{2,18} = 2.800$, $p = 0.087$, $\eta^2 = 0.237$).

Physiological, Flight Technical, and Subjective. Additional variables were tested for normality using the Kolmogorov-Smirnov (KS) test. For the variables that possessed a normal distribution, a General Linear Model ANOVA was used to test for significant differences between groups. If the normality assumption was violated, and no suitable transformation could be found, then a non-parametric Kruskal-Wallis test was used to test if a significant difference existed. Table 3 shows a summary of the results of the normality and significance tests. The significant treatments ($\alpha \leq 0.05$) are represented with bold, italicized text.

Table 3. Summary of Analysis Results

Condition	Dependent Variable	Normality Test	Significance Test	Simulator vs Jet	With vs Without NT	Maneuver Level
All Levels	Subjective Workload	KS = 0.06 p > 0.15	ANOVA	***$F_{1,106} = 5.73$*** ***$p = 0.018$***	***$F_{1,106} = 172.18$*** ***$p < 0.001$***	***$F_{2,106} = 31.14$*** ***$p < 0.001$***
	Situation Awareness	KS = 0.07 p = 0.14	ANOVA	***$F_{1,106} = 5.82$*** ***$p = 0.018$***	***$F_{1,106} = 147.64$*** ***$p < 0.001$***	***$F_{2,106} = 22.17$*** ***$p < 0.001$***
Low & Medium	Altitude Error RMS (ft)	KS = 0.06 p > 0.15	ANOVA	$F_{1,62} = 0.01$ $p = 0.91$	***$F_{1,62} = 22.59$*** ***$p < 0.001$***	***$F_{1,62} = 6.25$,*** ***$p = 0.015$***
	Roll Error RMS (deg)	KS = 0.13 p < 0.01	Kruskal-Wallis	H = 1.85 p = 0.17	***H = 22.21*** ***$p < 0.001$***	***H = 4.75*** ***$p = 0.029$***
High	Vertical Speed Error RMS (ft/min)	KS = 0.13 p = 0.10	Kruskal-Wallis	H = 1.51 p = 0.22	***H = 12.38*** ***$p < 0.001$***	N/A
All Levels w/ NT	Upper Display Intersection (%)	KS=0.098 p > 0.15	ANOVA	$F_{1,47} = 3.40$ $p = 0.072$	N/A	$F_{2,47} = 0.53$ $p = 0.59$
	Lower Display Intersection (%)	KS = 0.11 p = 0.06	Kruskal-Wallis	***H = 3.81*** ***$p = 0.05$***	N/A	H = 1.80 p = 0.41

The significant results for the NT, including the post-hoc Tukey pairwise t-test results, indicate that workload, SA, and error rates were higher when the NT task was present. A similar result was found with the maneuver's level of difficulty. The test apparatus showed evidence of being significantly higher in terms of workload ($F_{1,106} = 5.73$, $p = 0.018$). Figure 7 shows side-by-side boxplots of the breakdown between the test apparatus, the maneuver's level of difficulty, and the presence or absence of the NT task for both the subjective and CPC workload. The graphs show different stories, with the CPC results consistently higher than the self-assessed Bedford ratings. The CPC results are consistent with the results demonstrated by the participants' NT performance. It is possible the pilots were unable to properly assess their overall performance on the combined flight technical and NT task loading or that the task of selecting an appropriate Bedford rating was an additional burden on already overloaded cognitive resources.

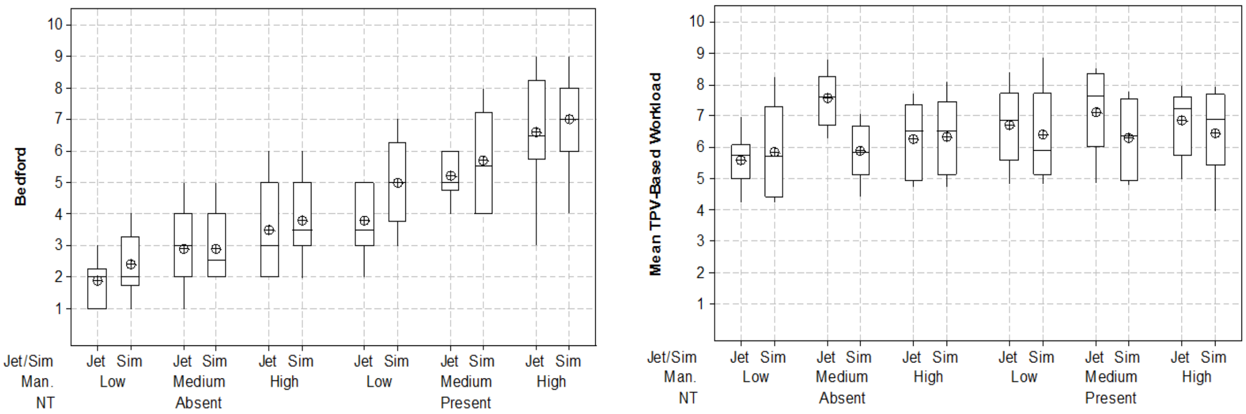

Figure 7. Side-by-Side Comparison of Workload Ratings

The eye tracking metrics were calculated as the percentage of time the intersection (the word intersection refers to the gaze vector intersecting with the display) occurred in a specified primary area of interest, which was designated to be the upper and lower displays for the initial study. Figure 8 presents the distributions of the Upper and Lower Intersections for the "NT present" condition. Analysis showed the Upper Intersection was nearly zero across all participants when the NT task was absent, which was expected since the upper display was solely used for the NT task. The graph does not include the percentage of time when their eye gaze moved beyond the primary area of interest.

The ANOVA for Upper Intersection indicated there was no significant difference ($F_{2,47} = 0.53$, $p = 0.59$) for the maneuver's level of difficulty. The same result was shown for the Lower Display Intersection ($H = 1.80$, $p = 0.41$). The EPs were asked to create a "scan pattern", which divided their attention between the NT task and the flight instruments, and is the likely cause of this result. There was evidence of differences between the simulation and live flight conditions. The Upper Display Intersection was almost significant ($F_{1,47} = 3.40$, $p = 0.072$), and the Lower Display Intersection showed evidence of being significant ($H = 3.81$, $p = 0.05$). These results indicate that the EP focused on the interior of the aircraft more during the simulator runs, likely due to fewer visual cues and external distractions.

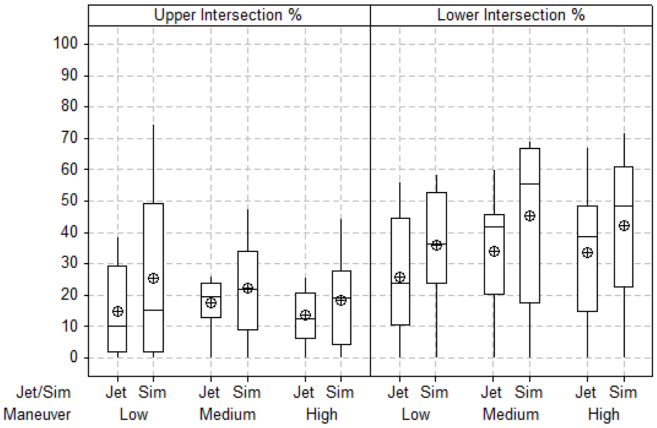

Figure 8. Percentage of Eye Gaze Intersection by Display (NT Present)

DISCUSSION

The primary goal of this study was to establish our experimental approach for objectively assessing cognitive state and training effectiveness. The study also gathered initial findings conducted, due to budgetary constraints, with a small sample size of low time pilots to lay the foundation for a larger follow-on effort. The results show that both the perceptual-cognitive and physiological assessment techniques can successfully differentiate between the test apparatus configurations and the maneuver's level of difficulty.

The workload evaluation using the perceptual-cognitive methodology demonstrated that flying is a cognitively demanding task regardless of the nature of the apparatus (simulator or jet). This result was observed by a drastic decrease (average of ~97%) in the NT displacement thresholds when comparing the NT scores at rest and the NT scores when the EPs were flying. Additionally, there was a significant decrease in flight technical performance when the NT task was present compared to when the NT was absent. This result is not surprising as a decrease in task performance is a common phenomenon when two tasks must be executed simultaneously.

The low displacement threshold values while flying suggest that the multi-object tracking component of the NT task was only slightly activated during the exercises, given that the values indicate the targets were barely moving. It is likely that the EPs were relying solely on the working memory component of the task. The hypothesized mechanism is as follows: 1) memorize the position of the indexed targets when presented, 2) maintain the slightly moving targets' positions in memory while focusing on flying, and 3) recall the position of the indexed targets from step 1. While the MOT was more or less deactivated from the NT task, it was still sensitive to the EPs cognitive workload.

The results demonstrated that both NT and physiological performance were more affected during live flight in the jet compared to the AIL simulator configuration. Consequently, it can be suggested that the cognitive load is significantly heavier in live flight. These results agree with the idea that there is additional physiological and environmental noise in real-world conditions compared to a laboratory context (Callan et al., 2015) and that brain dynamics differ in real-world environments compared to those of a laboratory (Mcdowell et al., 2013). Additionally, the measures were sensitive to the maneuver level of difficulty.

NEXT STEPS

This effort has confirmed that the perceptual-cognitive and physiological measures employed in the study are appropriate markers of cognitive workload and can be adaptable to complex environments, such as airplane cockpits. Decoding the cognitive state of pilots across various flight conditions will improve the quality of simulation-based training devices and training effectiveness. To achieve the long-term goal of this research initiative, the methods from this study will be employed in a subsequent effort with a larger and more diverse population, including subjects representing the skilled and expert levels that were not present in the current sample. Collecting and analyzing data from this broad pool will enable characterization of the physiological responses and perceptual-cognitive capabilities of individuals ranging from novice trainees to highly skilled operators in a range of training contexts. Additional follow-on efforts will consider longitudinal evaluation of subjects to classify the physiological evolution associated with skill acquisition.

Two modifications to the experimental approach were identified for the follow-on studies. First, the low level of NT performance indicates that the EPs are only using the working memory component of the NT task. To leverage the MOT component, the task difficulty will need to be altered, likely by reducing the number of targets to track from four to two. Second, the EPs reported some difficulty interacting with the NT software through the touchscreen interface. Limitations in the resolution and accuracy of the device resulted in selection of unintended targets when two spheres stopped near each other. The EPs had to deselect the incorrectly registered target and attempt to select the intended target. We will use verbal selection of targets as an alternative interaction method.

This objective assessment method, once refined, has broad applications within the simulation and training industry. This method will be used to develop intelligent semi-automated forces that behave in a more human-like manner and can dynamically respond to trainee state. Real-time assessment of cognitive state will also provide a means for personalizing learning. Personalization will be achieved through dynamically adaptive content that classifies and manages cognitive engagement during a training exercise. For example, if the state assessment system identifies that a trainee is cognitively overloaded and has poor task performance, the exercise can quickly be simplified to allow the trainee to still receive benefit from the event. Conversely, if the state assessment system identifies that a trainee is under-tasked and is performing at a high level (a skill has been mastered), the complexity of the exercise can be increased or the trainee can be accelerated through the curriculum to more advanced tasks.

REFERENCES

Callan, D.E., Durantin, G., & Terzibas, C. (2015). Classification of single-trial auditory events using dry-wireless EEG during real and motion simulated flight. *Frontiers in Systems Neuroscience*, *9*:11.

Engler, J., Schnell, T., & Walwanis, M. (2013). Deterministically Nonlinear Dynamical Classification of Cognitive Workload. *Proceedings of the Interservice/Industry Training, Simulation and Education Conference.* Arlington, VA: National Training and Simulation Association.

Faubert, J. (2013). Professional athletes have extraordinary skills for rapidly learning complex and neutral dynamic visual scenes. *Scientific reports*, *3*, 1154.

Faubert, J., & Sidebottom, L. (2012). Perceptual-Cognitive Training of Athletes. *Journal of Clinical Sports Psychology*, *6*(1), 85-102.

Govindan, R.B., Narayanan, K., & Gopinathan, M.S. (1998). On the evidence of deterministic chaos in ECG: Surrogate and predictability analysis. *Chaos*, *8*(2), 495 – 502.

Harenberg, S., McCaffrey, R., Butz, M., Post, D., Howlett, J., Dorsch, K. D., & Lyster, K. (2016). Can multiple object tracking predict laparoscopic surgical skills? *Journal of Surgical Education*, *73*(3), 386–390.

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

Kozma, R., & Freeman, W.J. (2002). Classification of EEG patterns using nonlinear dynamics and identifying chaotic phase transitions. *Neurocomputing*, *44*, 1107 – 1112.

Legault, I., & Faubert, J. (2012). Perceptual-cognitive training improves biological motion perception: evidence for transferability of training in healthy aging. *Neuroreport*, *23*(8), 469 – 473.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2), Suppl 2:467+.

Mangine, G. T., J. R. Hoffman, A. J. Wells, A. M. Gonzalez, J. P. Rogowski, J. R. Townsend, A. R. Jajtner, K. S. Beyer, J. D. Bohner, G. J. Pruna, M. S. Fragala & J. R. Stout (2014). Visual Tracking Speed Is Related to Basketball-Specific Measures of Performance in NBA Players. *Journal of Strength and Conditioning Research*, *28*(9), 2406-2414.

Mcdowell, K.T., Lin, C.T., Oie, K.S., Jung, T.P, Gordon, S., Whitaker, K.W., Li, Y.S., Lu, S.W., & Hairston, W.D. (2013). Real-World Neuroimaging Technologies. *IEEE Access*, *1*, 131-149.

Owis, M. I., Abou-Zied, A. H., Youssef, A. B., & Kadah, Y. M. (2002). Study of features based on nonlinear dynamical modeling in ECG arrhythmia detection and classification. *IEEE transactions on Biomedical Engineering*, *49*(7), 733-736.

Parsons, B., Magill, T., Boucher, A., Zhang, M., Zogbo, K., Bérubé, S., Scheffer, O., Beauregard, M, & Faubert, J. Enhancing Cognitive Function Using Perceptual-Cognitive Training. *Clinical EEG and Neuroscience, 47* (1), 1–11.

Richter, M., & Schreiber, T. (1998). Phase space embedding of electrocardiograms. *Physical Review E*, *58*(5), 6392.

Romeas, T., Guldner, A., & Faubert, J. (2016). 3D-Multiple Object Tracking training task improves passing decision-making accuracy in soccer players. *Psychology of Sport and Exercise*, *22*, 1-9.

Roscoe, A.H. (1984). Assessing pilot workload in flight. In *AGARD Conference Proceedings, No. 373 – Flight Test Techniques*, 1032.

Schnell, T., & Engler, J. (2014). Entropic skill assessment of unmanned aerial systems (UAS) operators. *Journal of Unmanned Vehicle Systems*, *2*(02), 53-68.

Schreiber, B.T., Bennett, W., Jr., & Stock, W.A. (2006). Distributed mission operations within-simulator training effectiveness baseline study: Metric development and objectively quantifying the degree of learning. (Report No. AFRL-HE-AZ-TR-2006-0015, Vol. 2). Mesa, AZ: Warfighter Readiness Research Division.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257-285.

Taylor, R.M. (1989). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations, CP478*.

United States General Accounting Office. (2013). Army and Marine Corps Training: Better performance and cost data neeed to more fully assess simulation-based efforts. Report to Congressional Committees. (Report GAO-13-698). United States Government Accountability Office, Washington, D.C.

Vartanian, O., Coady, L., & Blackler, K. (2016). 3D multiple object tracking boosts working memory span: Implications for cognitive training in military populations. *Military Psychology*, *28*(5), 353-360.