USAARL-TECH-BB--2022-13

UNITED STATES ARMY AEROMEDICAL RESEARCH LABORATORY

Annotated Bibliography: Cognitive Workload of Aviators With Various Levels of Experience

Jon Vogl, Cristina Delgado-Howard, Aaron McAtee, Christopher Aura, Amanda Hayes, & Paul St. Onge

Human Subjects Use Statement

The U.S. Army Aeromedical Research Laboratory Determination Official determined that this activity does not constitute research as defined under the human subjects protection regulations, as it is not “a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge.” [32 CFR 219.102(d)].

REPORT DOCUMENTATION PAGE			*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
19-11-2021	Bibliography	

4. TITLE AND SUBTITLE
Annotated Bibliography: Cognitive Workload of Aviators with Various Levels of Experience

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Vogl, J.[1,2], Plummer, H.,[1,3], Delgado-Howard, C.,[1,3]; McAtee, A.[1,2], Aura, C.[1], Hayes, A.[1], & St. Onge, P.[1]

5d. PROJECT NUMBER
6.2

5e. TASK NUMBER
2100

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Aeromedical Research Laboratory
P.O. Box 620577
Fort Rucker, AL 36362

8. PERFORMING ORGANIZATION REPORT NUMBER
USAARL-TECH-BB--2022-13

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Development Command
504 Scot Street
Fort Detrick, MD 21702-5202

10. SPONSOR/MONITOR'S ACRONYM(S)
USAMRDC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
[1]U.S. Army Aeromedical Research Laboratory; [2]Goldbelt Frontier, LLC; [3]Oak Ridge Institute for Science and Education

14. ABSTRACT
This annotated bibliography serves as a reference document for an upcoming systematic literature review on composite cognitive workload assessment in the aviation domain. In this review, studies focusing on testing experienced pilots working in aviation platforms (simulators and real aircraft) were analyzed. This focus was chosen in an effort to control for individual difference and task demand factors that can affect cognitive workload across studies (ideally reducing the variance expected from untrained subjects engaging in novel tasks). For each reviewed article, the experience levels of pilots (as defined by reported flight hours) and the types of cognitive workload assessment measures used were used to categorize articles. In the resulting systematic literature review, continued analysis examining the response profiles in terms of the AIDs taxonomy of cognitive workload assessment will be reported, among other findings. A brief description of the methodology used to derive the final list of articles and the list of categorized article summaries follows.

15. SUBJECT TERMS
cognitive workload, expert, experienced, novice, AIDs taxonomy

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Loraine St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	35	**19b. TELEPHONE NUMBER** *(Include area code)* 334-255-6906

This page is intentionally blank.

Summary

This annotated bibliography was compiled from a comprehensive systematic literature review on the topic of composite cognitive workload assessment in the aviation domain. In this review, studies focusing on testing experienced pilots working in aviation platforms (simulators and real aircraft) were collected and analyzed. This focus was chosen in an effort to control for individual difference and task demand factors that can affect cognitive workload across studies (ideally reducing the variance expected from untrained subjects engaging in novel tasks). For each reviewed article, the experience levels of pilots (as defined by reported flight hours) and the types of cognitive workload assessment measures used were used to categorize articles. In the resulting systematic literature review, continued analysis examining the response profiles in terms of the associations, insensitivities, and dissociations (AIDs) taxonomy of cognitive workload assessment will be reported, among other findings. A brief description of the methodology used to derive the final list of articles and a list of categorized article summaries follows.

This page is intentionally blank.

Acknowledgements

This research was supported in part by an appointment to the Postgraduate Research Program at the U.S. Army Aeromedical Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Army Medical Research and Development Command.

This page is intentionally blank.

Table of Contents

List of Tables

List of Figures

This page is intentionally blank.

Introduction

In recent decades, the unit of work has seen a shift from the joule to the byte. No longer singularly concerned with how systems challenge an operator's physical abilities, research shifted focus to an operator's cognitive abilities. Emerging and future systems promise additional multi-tasking and information management work, potentially inducing cognitive workload beyond a typical operator's ability. This progression is happening across domains from manufacturing to systems management, and transportation; Army aviation is no different.

Future Vertical Lift aircraft will have advanced flight decks to provide heightened levels of situation awareness to secure a tactical advantage and will challenge the limits of human cognitive ability. These advanced flight decks offer pilots multiple video feeds from other aircraft, ground stations, terrain maps, displays from external sensors on the aircraft, force tracking devices, and communication devices to increase tactical advantages. However, the benefits of heightened situation awareness will have to be balanced with the increased likelihood of cognitive overload and the systems available to assist with information management.

To aid pilots working with these advanced systems, operator state monitoring (OSM) systems act as a way to identify adverse cognitive states and initiate corrective action. OSM systems capture a variety of biological markers known to correlate with cognitive activity using physiological sensors, such as electrocardiogram (ECG), electroencephalogram (EEG), eye tracking systems, respiration sensors, temperature sensors, and blood oxygenation sensors, among others. Much like any mechanical component of a system, signals of various physiological phenomena, such as cognitive workload, injury, fear, or fatigue, can be processed in real-time to provide a constant read-out of the state of the human factor of the system. Additionally, OSM systems can use performance data and subjective reports from pilots to influence cognitive state identification. Mission commanders or artificially intelligent systems can use this information to aid a pilot or mitigate a mishap.

This review examines the use of OSM systems that assess cognitive workload states through the use of combined cognitive workload metrics. The moment-to-moment physiological snapshot provided by standard OSM systems is only part of the picture in terms of cognitive workload assessment. Other facets of cognitive workload assessment include changes in performance measures when workload levels are too low or too high and changes in the subjective responses of pilots experiencing cognitive workload. These three assessment techniques (physiological, performance, and subjective measures) can be combined to yield a composite cognitive workload measure. However, it is not uncommon to see cognitive workload measures dissociating across these techniques even within the same individual (Hancock & Matthews, 2019). To review this space in the literature, a systematic literature review of aviation studies focused on cognitive workload assessment spanning the past decade (2010-2020) has been conducted. This report serves as an overview of the cognitive workload assessment framework used to structure the literature review and presents an organized annotated bibliography of the studies included in the aforementioned literature review.

Brief Summary of Cognitive Workload and Its Assessment

Since its introduction to the literature in the 1960s, cognitive workload has been represented by various definitions, terminology, and frameworks (Westbrook, Anderson, & Pietrzak, 1966). With other concepts, such as stress, fatigue, and different types of workload, muddying the jargon of cognitive workload, De Waard (1996) attempted to focus the vocabulary used when discussing cognitive workload and its subcomponents. In Cain's (2007) review, a list of contemporary definitions of cognitive workload found in the literature was presented. An examination of the literature shows a common resource-demand model has taken a dominant position in the understanding of cognitive workload. The resource-demand framework simply states that when cognitive demand surpasses cognitive resource availability, cognitive overload is experienced. In 2018, a concept analysis of cognitive workload was performed by Van Acker et al. to derive a formal definition of cognitive workload that parallels the resource-demand models. The definition presented by Van Acker et al. (2018) verbatim:

> *"Mental workload is a subjectively experienced physiological processing state, revealing the interplay between one's limited and multidimensional cognitive resources and the cognitive work demands being exposed to."*

This review uses the Van Acker et al. (2018) definition as the foundational understanding of cognitive workload. Cognitive workload exists due to the interaction between a human operator and a specific task (in a given contextual environment). Cognitive workload and the resulting induced behavior is defined by parameters of individual differences and task-defined resource demand. Operators with larger working memories, specific personalities, propensities for anxiety, higher emotional intelligence, and different cognitive styles experience different levels of cognitive overload effects and utilize different workload management techniques. Additional, more malleable day-by-day, individual-level drivers of cognitive overload perception include operator-level or mission-level set criteria, motivation, strategies, sleep deprivation, and experience. On the task side of the equation, each task performed by an operator draws on or demands resources from the operator. Task analyses, using multidimensional attentional resource frameworks such as Multiple Resource Theory (Wickens, 1984), can pinpoint the demand placed on each resource available from the operator. Individual tasks are not performed in a vacuum in an applied setting. The interaction between two tasks demanding resources from an operator makes the multi-tasking scenario even more demanding than the summation of the two tasks performed individually in succession. This interaction is referred to as task interference (Wickens, 2008). Controlling the individual and task-level aspects of the cognitive workload equation allows for more reliable analyses of effects and behaviors that vary as a function of workload.

The Van Acker et al. (2018) definition also states that cognitive workload is a phenomenon that has a subjective experience tied to changes in physiological processing. Cognitive workload assessment has classically tapped into this experience by utilizing three primary methods: Performance (primary task and secondary task), physiological, and subjective measures. Each of these measures has strengths and weaknesses across varying measurement criteria, as defined by O'Donnell and Eggemeier (1986). The performance of an operator on a primary or secondary task acts as the most direct measure of cognitive workload. Primary task performance measures follow an inverse relationship with cognitive workload levels. However, primary task performance measures are generally not sensitive to changes in cognitive demand until cognitive overload occurs. Secondary task performance measures serve as measures of spare capacity (i.e., subsidiary task paradigm) or as a method to load on specific cognitive resources (i.e., loading task paradigm) (O'Donnell & Eggemeier, 1986). These methods are generally reserved for laboratory assessment as they cause higher levels of intrusion on the primary task. Physiological measures of cognitive workload are used to monitor the physiological changes that occur with increasing cognitive activity and arousal. Various physiological metrics are employed for cognitive workload assessment, each with their own typical patterns of response. For example, pupillometry yields increased pupil dilation under cognitive load, ECG reveals a reduction of heartrate variability under cognitive load, and respiration rate increases under cognitive load. Unfortunately, due to physiological signals exhibiting the summation of the human arousal experience, it is difficult to identify each unique factor affecting the signal. Lastly, subjective measures are a common workload assessment technique that query the operator on their perceived experience of cognitive workload with a specific task or task sequence. Some commonly utilized subjective measures of cognitive workload include the NASA Task Load Index (NASA TLX), Workload Profile (WP), Modified Cooper-Harper (MCH), Crew Status Survey (CSS), and Rating Scale Mental Effort (RSME) (Vogl et al., 2020). Subjective measures typically follow task performance, as they otherwise heavily intrude on primary task performance. Each of these types of cognitive workload assessment tap into the same experience for the operator and task being evaluated; however, the stories told by each technique may not always be similar.

Composite Cognitive Workload Assessment Framework

With each type of cognitive workload assessment technique having their own strengths and drawbacks, it seems natural to combine performance, physiological, and subjective measures into a composite measure of cognitive workload. The logic follows that with each of these cognitive workload responses being measured from the same individual that the responses would correlate with each other. When a composite measure shows increasing cognitive workload in each individual measure, one could be confident that the studied operators were experiencing higher levels of cognitive workload or vice versa with matching decreasing cognitive workload responses. In the event that different responses come from each cognitive workload assessment metric, the experience of the studied operator becomes less clear and more puzzling. For example, an operator can report low levels of cognitive workload in subjective measures but their physiological measures indicate increasing levels of workload while their performance metrics remain stable. Likewise, the same inconsistencies can be modelled across the different cognitive workload measures each with responses that indicate high, low, or stable cognitive workload. Hancock and Matthews (2019) explored the concept of associations, insensitivities, and dissociations (AIDs) of cognitive workload assessment to create a framework in which to understand the possible states of composite workload assessment metrics. The matrix defining these possible composite cognitive workload assessment states can be seen in Figure 1.

This space is intentionally blank.

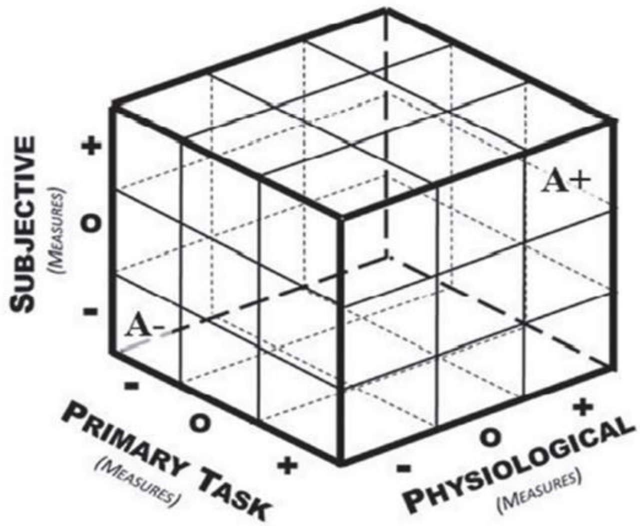

Figure 1. The matrix of Association, Insensitivity, a Dissociation (AIDs) framework of cognitive workload assessment techniques by Hancock and Matthews (2019). Each measure can indicate an increasing (+), decreasing (-), or stable (o) cognitive workload response. With each state represented by a cube within the matrix, 27 combinations of outcomes across performance (primary task), physiological, and subjective measures are possible. When measures agree with each other (i.e., all measures are either showing a decrease or increase in cognitive workload), a double association (denoted by A- for decreasing and A+ for increasing measures) occurs.

The Hancock and Matthews (2019) AIDs taxonomy for composite cognitive workload assessment states presents the three primary methods of cognitive workload assessment along the axes of a cube-shaped matrix. Each method allows for one of three responses: increasing cognitive workload response (+), decreasing cognitive workload response (-), and stable (i.e. insensitive) cognitive workload response (o). Combining each individual measure's response outcome yields a three dimensional matrix defining 27 unique states of a composite cognitive workload measure. When responses from different types of workload measures match one another (e.g., increasing workload as defined by both physiological and subjective measures), an association occurs between the two measures. In the event that two measures' responses disagree with each other, a dissociation occurs. A double association (as denoted by the A+ and A- states in Figure 1) occurs in the event that all three measures report the same responses (i.e., all measures show a matched response of increasing, stable, or decreasing cognitive workload). Likewise, double dissociations occur when then all measures disagree with each other. While double associations simplify the problem of interpretation of cognitive workload data, recognizing factors that affect the convergence of measures can aid in understanding why dissociations occur. Hancock and Mathews (2019) elaborate on these issues of convergence between measures and identify common problems that can influence mismatching responses between measures. Factors such as the granularity between measures, the timing of cognitive workload responses across measures, self-regulating strategies, and workload history can all lead to inconsistencies between different measurement techniques. Ultimately, these problems are still unsolved, but recognizing their presence can aid in the interpretation of even the most dissociated data sets.

Literature Review Framework

This annotated bibliography emerged by conducting a systematic literature review on composite cognitive workload assessment in the aviation domain. In this review, studies focusing on testing experienced pilots working in aviation platforms (simulators and real aircraft) were analyzed. This focus was outlined in an effort to control for individual difference and task demand factors that can affect cognitive workload across studies, ideally reducing the variance expected from untrained subjects engaging in novel tasks. For each reviewed article, the experience levels of pilots (as defined by reported flight hours) and the types of cognitive workload assessment measures used were used to categorize articles. In the resulting systematic literature review, continued analysis examining the response profiles in terms of the AIDs taxonomy of cognitive workload assessment will be reported, among other findings. A brief description of the methodology used to derive the final list of articles and a list of categorized article summaries follows.

Methods

A systematic literature review was conducted to explore the literature associated with cognitive workload measurement techniques. Initial searches were conducted across three databases: Psych Info, ProQuest, and Web of Science. All searches were limited to a 10-year period (2010-2020) to focus on the state-of-the-art in cognitive workload assessment. Searches were also limited to the English language, human subject use, and a cognitive workload focus. Each search included a Boolean list of terms to refer to the use of trained, professional populations with a special emphasis on military and aviation populations used in the papers retrieved. By controlling for experience in the experimental task, it is thought that cognitive workload assessment will be more consistent, as opposed to novice populations performing novel tasks. The final list of around 1200 articles spanned multiple domains of operation and provided counts of the prevalence of cognitive workload research in each domain (to be reported in an upcoming review paper). For this review, our focus is the aviation domain and studies that utilize trained pilots in simulation or real aircraft settings.

Fifty-nine aviation-related cognitive workload articles were reviewed between January 1 and March 1 of 2020. Table 1 summarizes the categorization and article counts. The analyzed articles were first categorized by the experience level of the subjects participating in the study. Articles that did not report the subject's experience in flight hours were excluded from this review. The remaining articles were sorted by the reported experience in flight hours and categorized into three levels: Novice, intermediate, and expert. The threshold for each category was defined by the flight hour licensure requirements set forth by Code of Federal Regulations (CFR). The novice category included pilots who had less than 150 flight hours of experience participating in rotary aircraft studies and 250 flight hours for fixed wing aircraft. The ceiling of this threshold marks the transition from private pilot status (requiring 40 flight hours, as described in CFR Title 14, Chapter I, Subchapter D, Part 61, Subpart E, Section 61.109) to commercial pilot status. The intermediate category consists of pilots with over 150 flight hours who would be eligible for a commercial pilot's license (requiring 150 flight hours for rotary aircraft and 250 flight hours for fixed wing aircraft, as described in CFR Title 14, Chapter I, Subchapter D, Part 61, Subpart F, Section 61.129). Subjects who reported having over 1200 flight hours for rotary aircraft and 1500 hours for fixed wing aircraft (the requirements for an airline transport pilot, as described in CRF Title 14, Chapter 1, Subchapter D, Part 61, Subpart G, Sections 61.159 and 61.161) in aviation-related research studies were categorized into the expert group.

The second categorization process consisted of sorting articles by the types of cognitive workload assessment techniques utilized in the study. The three primary categories include performance measures, physiological measures, and subjective measures. Articles within each experience level were sorted by the number of cognitive workload measurement types utilized in the study, with studies that used all three assessment types being listed first, descending down to studies that only used a single type of cognitive workload assessment. The counts of assessment type and flight experience level are summarized in Table 2. Article summaries were written in a way that capitalized on reporting the results of each type of measure to facilitate the analysis in the AIDs of cognitive workload assessment framework. Additional analyses of the findings among the articles in this review will be reported in a forthcoming formal systematic literature review paper.

Table 1. Summary of Articles Based on Exclusion Criteria

Experience Level	Description	Number of Articles
Excluded	Articles that did not report mean flight hours	23
Novice	• Less than 150 Flight Hours—Rotary-wing • Less than 250 Flight Hours—Fixed-wing	6
Intermediate	• Between 150 and 1199 Flight Hours—Rotary-wing • Between 250 and 1499 Flight Hours—Fixed-wing	11
Expert	• 1200+ Flight Hours—Rotary-wing • 1500+ Flight Hours—Fixed-wing	19
Total Number of Articles Reviewed		**59**

This space is intentionally blank.

Table 2. Summary of Articles Based on Assessment Types

Experience Level	**3 Assessment Types Present**	**2 Assessment Types Present**	**1 Assessment Type Present**
Novice	2	2	2
Intermediate	4	4	3
Expert	3	10	6
Total	**9**	**16**	**11**

This space is intentionally blank.

Annotated Bibliography

Expert Population

All Assessment Types—Performance, Physiological, and Subjective.

Dahlstrom, N., Nahlinder, S., Wilson, G. F., & Svensson, E. (2011). Recording of psychophysiological data during aerobatic training. ***The International Journal of Aviation Psychology*****, 21, 105-122.**

Search Terms: ECG, EOG, EEG, Mental Workload, Aviation

Subjects (n = 7) performed aerobatics flight sequences as part of flight school training. ECG, EEG, electrooculography (EOG) data were collected during flight. Questionnaires were administered to assess subjective mental workload, the success of their performance, and the experienced task difficulty during each flight segment and for the individual maneuvers in each sequence. Post flight the pilots rated the outcome of the flight and the factors that may have contributed to their performance. An analysis of variance (ANOVA) with flight segment as a repeated measurement showed significant effects of flight segments, $F(6, 30) = 3.55$, $p < .01$. Ratings of mental workload were higher for the aerobatics sequences than for other flight segments, with landing being the highest thereafter. Ratings of performance and difficulty display a similar pattern, $F(6, 30) = 3.27$, $p < .05$, and $F(6, 30) = 5.18$, $p < .001$, respectively, with relatively lower performance ratings and higher difficulty ratings for the aerobatics sequences. The results for ratings of mental workload and difficulty both show significant main effect of maneuver: mental workload, $F(5, 25) = 11.09$, $p < .0005$; difficulty, $F(5, 25) = 8.622$, $p < .0005$, but no main effect of sequence, nor of the interaction (Maneuver × Sequence). Post-hoc analyses reveal that “wing-over” has significantly lower ratings of mental workload and difficulty than the other maneuvers (except the loop). The Cuban eight and Immelman have higher ratings of mental workload and difficulty than the other maneuvers (except the hammerhead).

Nittala, S., Elkin, C., Kiker, J., Meyer, R., Curro, J., Reiter, A., Xu, K., & Devabhaktuni, V. (2018). Pilot skill level and workload prediction for sliding-scale autonomy. ***2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*****, 1166-1173.**

Search Terms: Heart Rate, Task analysis, Autonomous Systems, Data Collection, Prediction Algorithms, Aircraft, Indexes

Subjects (n = 15; 12 novice pilots, 3 expert pilots) completed a series of flight simulator tasks to test the ability to predict skill level and workload using various machine learning approaches to aid sliding-scale autonomous assistive systems. The X-plane flight simulator was used for performance data collection, a PulseSensor ear clip connected to an Arduino collected heart rate (HR) data, and the NASA-TLX was used to collect subjective workload data (with a primary focus on the Mental Demand subscale). Reported results of the NASA-TLX Mental Demand subscale mean values show a significant difference between the novice and expert groups (no p value reported). Data for the predictive models were processed using logistic regression, SVM (Linear kernel), SVM (RBF kernel), random forest, and gradient-boosted trees algorithms. Prediction results show that prediction of pilot skill level was performed with near-perfect accuracy using only flight performance measures across algorithms, while workload prediction resulted in only moderate success when using flight performance or heart rate alone ($R2$ values around 0). Using a two-stage approach that utilized predicted skill levels, individual novice and expert regression models, flight data, and HR data, workload was more accurately predicted ($R2$ values around 0.3). Further analysis revealed more computationally complex algorithms yield better prediction of mental workload using the two-stage approach and SVM (RBF kernel) algorithm ($R2$ values around 0.59). This study offers a snapshot and review of how composite cognitive workload metrics can be utilized to obtain functional human-machine adaptive automation systems.

Yang, J. H., Kennedy, Q., Sullivan, J., & Fricker, R. (2013). Pilot performance: Assessing how scan patterns & navigational assessments vary by flight expertise. *Aviation, Space, and Environmental Medicine*, 84, 116-124.

Search Terms: Adult, Aviation, Eye Movements, Humans, Male, Professional Competence, Task Performance and Analysis

Subjects (n = 12) completed a helicopter overland navigation task that was divided into easy and difficult routes. Pilots were classified as expert of novice based on previous experience with overland navigation. To begin, they studied a map of the route and then they flew the route in a simulator. The route included 12 way points that pilots were told to stay within 0.5 kilometers (km) from in order to be considered on-track. Their visual scan patterns that included looking out-the-window and at a topographical map were assessed. Root mean square (RMS) error increased and navigation accuracy decreased from the easy route section to the difficult section [$t(11) = 5.171$, $P < 0.001$ and $t(11) = 3.924$, $P < 0.01$, respectively]. Ten pilots were on course for the easy route but only three pilots were on course for the difficult route. Gaze parameters were not different between routes. A negative correlation between navigation accuracy with median dwell and out-the-window (OTW) dwell (median dwell, $r = -0.45$, $P < 0.1$; median OTW dwell, $r = -0.52$, $P < 0.05$) during the easy route but no significance was observed in the difficult route. Navigation accuracy was correlated negatively with RMS error, out-the-window dwell duration, and post-task route assessment on the easy route ($r = -0.52$, $P < 0.05$; $r = -0.52$, $P < 0.05$; $r = -0.55$, $P < 0.05$).

This space is intentionally blank.

Combination of Two Assessment Types—Performance, Physiological, and Subjective.

Cheung, B., Craig, G., Steels, B., Sceviour, R., Cosman, V., Jennings, S., & Holst, P. (2015). In-flight study of helmet-mounted symbology system concepts in degraded visual environments. *Aerospace Medicine and Human Performance, 86*(8), 714-722.

Search Terms: Aerospace Medicine, Aircraft/Instrumentation, Canada, Computer Simulation, Confusion/Prevention and Control, Data Display, Head Protective Devices, Humans, Male, Military Personnel, Orientation, Space Perception, Spatial Navigation, Task Performance and Analysis

Subjects (n = 10) were Royal Canadian Air Force (RCAF) Griffon pilots participated in an in-flight study of a 3-dimensional conformal Helmet Display Tracking System (HDTS) and the BrownOut Symbology System (BOSS) aboard an advanced system research aircraft. The subjects performed a two-stage departure followed by a single-stage approach for each symbology system. Presentation of the two symbology systems was randomized. Subjective measurements were situation awareness, mental effort, perceived performance, perceptual cue rating, NASA TLX, and physiological response. Objective performance metrics were aircraft speed, altitude, attitude, distance from the landing point, control position, and control activity. Repeated measures ANOVAs and planned comparison tests for the subjective and objective responses were performed. During departure the HDTS provided better situation awareness ($p < 0.05$), required less mental effort ($p < 0.05$) based on the overall NASA TLX score and the Cooper Harper mental workload score, and resulted in better flight performance ($p < 0.05$) than the BOSS. In the approach phase, the HDTS required less mental effort ($p < 0.05$) and provided better situation awareness ($p \leq 0.05$), and resulted in better flight performance ($p < 0.05$) than the BOSS.

Cheung, B., McKinley, R. A., Steels, B., Sceviour, R., Cosman, V., & Holst, P. (2015). Simulator study of helmet-mounted symbology system concepts in degraded visual environments. *Aerospace Medicine and Human Performance, 86*, 588-598.

Search Terms: Aerospace Medicine/Instrumentation, Aircraft/Instrumentatio n, Computer Simulation, Confusion/Prevention and Control, Humans, Male, Spatial Navigation/Physiology

During approach and departure in rotary wing aircraft, a sudden loss of external visual reference precipitates spatial disorientation. There were 13 RCAF Griffon pilots who participated in a simulator study of 3 symbology sets used in a CH146 helicopter. They compared the following three symbology systems: CH146 AN/AVS Aviator's Night Vision Imaging System (AVS7), HDTS, and the BOSS. For each symbology system, pilots performed a two-stage departure followed by a single-stage approach. The presentation order of the symbology systems was counterbalanced across the pilots. Subjective measurements included situation awareness, mental effort, perceived performance, perceptual cue rating, NASA TLX, and physiological response. Objective performance included aircraft speed, altitude, attitude, distance from the landing point, control position, and control activity. Repeated measures ANOVAs and planned comparison tests for the subjective and objective responses were performed. During departure both HDTS and BOSS showed significant improvement over AVS7 ($p < 0.01$), i.e., increased situation awareness, lessened mental effort, and improved performance. For the NASA TLX, both HDTS and BOSS were rated significantly better ($p < 0.01$) in lessening mental demand, physical demand, temporal demand, performance, effort, frustration, and overall workload than the AVS7. HDTS was rated significantly better in lessening mental demand ($p < 0.05$), physical demand ($p < 0.02$), temporal demand ($p < 0.01$), effort ($p < 0.01$), and frustration level ($p < 0.04$) over BOSS. For perceptual cue rating (PCR), both HDTS and BOSS showed significant improvements ($p < 0.01$) in attitude and horizontal and vertical translational cues over the AVS7. For the approach, the HDTS system was rated to be significantly better in situation awareness when compared to BOSS ($p < 0.03$) and the AVS7 ($p < 0.01$). For the modified Cooper Harper rating on mental effort, HDTS was significantly better ($p < 0.05$) than the AVS7. HDTS performed significantly better than the AVS7 in temporal demand ($p < 0.02$), performance ($p < 0.01$), and frustration level ($p < 0.01$). For PCR, both HDTS and BOSS were significantly better than the AVS7 in all cueing categories ($p < 0.01$). Objective measures: HDTS and BOSS significantly better than the AVS7 ($p < 0.05$) for distance to initial position on departure and distance to designated landing point. HDTS significantly better than AVS7 on heading error ($p < 0.05$). Both HDTS and BOSS significantly better than AVS7 on departure heading root mean square error (RMSE), altitude RMSE, and approach heading RMSE ($p < 0.05$).

Diaz-Piedra, C., Rieiro, H., Suarez, J., Rios-Tejada, F., Catena, A., & Di Stasi, L. (2016). Fatigue in the military: Towards a fatigue detection test based on the saccadic velocity. *Physiological Measures, 37*(9), N62-N75.

Search Terms: Adult, Fatigue/Diagnosis/Physiopathology, Humans, Male, Military Personnel, Pilots, Saccades

The aim of this study was to assess inflight pilot workload using both physiological and multidimensional subjective ratings measurements, heart rate, and NASA TLX, respectively. Ten male pilots from China Airline Co. LTD, Taipei, Taiwan participated. Mean age of participants was 44.5 years with a range between 28 to 58. Mean flying experience was 9025 hours, with a range of 1000 to 25000 hours. All subjects were qualified to fly Boeing 747-400 aircraft. Study conducted in a Boeing 747-400 flight simulator. The sequence of the simulated tasks for all participants was: takeoff from runway 05L at Chiang Kai Shek international airport, turn left to a heading of 230° on attaining an altitude of 4000 feet (ft), descend to 2500 ft, turn left to a heading of 080°, then approach and land on the same runway using the instrument landing system (ILS). Flight time varied from 25-30 minutes. ECG conducted throughout the test using a portable Cardiovis ECG system. Mean HR and incremental HR (ΔHR) were indices of physiological workload. Resting HR was collected after experiment to serve as a baseline. An index of heart rate variability (HRV) was calculated. Peak HR observed during takeoff (83.2 beats per minute [bpm]) and landing (88.6 bpm). Max incremental HR observed during takeoff (14.2 bpm) and landing (18.8 bpm). A NASA TLX overall workload score was calculated using a weighted average of the six workload dimensions. Weighted score was different for each subject, depending on which dimensions each participant identified as most important for determination of workload. An overall workload score was calculated for each participant for each flight phase (takeoff, cruise, approach, and landing). Significant differences in ΔHR across flight phases, $F(3, 36) = 18.6$, $p < 0.001$. Mean ΔHR was greatest during landing (18.8 bpm), followed by takeoff (14.2 bpm), approach (10.6 bpm), and cruise (7.1 bpm). Significant difference in index of HRV between the four flight phases, $F(3, 36) = 21.6$, $p < 0.001$. Index of HRV lowest for approach (13.4 milliseconds [ms]), compared to landing (27.1 ms), takeoff (28.8 ms), and cruise (36.3 ms). Significant differences were found in overall NASA TLX scores between the four flight phases, $F(3, 36) = 11.1$, $p < 0.001$. Overall NASA TLX scores significantly higher for landing (74.6 ± 10.1), takeoff (70.7 ± 9.4), and approach (68.1 ± 8.7), in comparison to cruise (52.6 ± 10.1). NASA TLX results indicate that the main contributors to workload were mental, temporal, and performance demands during the takeoff phase; mental, performance, and physical during cruise; mental, performance, and temporal during the approach phase; and mental, performance, and effort during the landing phase. There was significant correlation between ΔHR and TLX. Correlation coefficient was 0.81 ($n = 40$, $p < 0.001$) and the calculated linear regression equation was NASA TLX = 52.16 + 1.12 * ΔHR.

Eichinger, A. & Kellerer, J. (2014). Between laboratory and simulator: A cognitive approach to evaluating cockpit interfaces by manipulating informatory context. *Cognition, Technology & Work, 16,* 417-427.

Search Terms: Jet Aircraft, Interface Evaluation, Informatory Context, Panoramic Display, Trackball, Touch Screen, Correspondence-Driven Domain

Subjects completed two different aiming tasks (single static targets versus targets that appear both static and moving with an additional moving distractor symbol) by using a mouse or touchscreen to select the targets as quickly as possible. Two types of control devices were compared under three qualities of informatory load and operationalized through three different additional tasks that had two levels of difficulty (easy and difficult). The additional tasks consisted of visual, cognitive, and motor tasks. For the additional visual task, subjects had to identify a single white circle as the target among 50 similarly shaped and colored circles. Subjects had to select the circle by using the left or right arrow key on a keyboard. For the cognitive task, recorded numbers between 1-9 were shown and then a single target number was presented. Subjects had to verbalize whether the number was part of the preceding sequence of numbers. The easy condition had a sequence of five numbers and the difficult condition had sequences of eight numbers. For the motor task, a wooden stencil was placed on a keyboard to record when a stylus was plugged into the circular slots. In the easy condition, the slot diameter was 25 millimeters (mm) and the stylus was 23 mm; in the difficult condition, the slots were 8 mm and the stylus was 7 mm. Subjects had to place the stylus in the pegboard slots in a clockwise order as quickly and accurately as possible. The results indicated that touch interaction had greater aiming performance than trackball interaction and difficult additional tasks result in high workload and lower task performance.

Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., & Wickens, C. (2001). Comparison of expert and novice scan behaviors during VFR flight. *Engineering Psychology and Cognitive Ergonomics*. HCI 2016 International.

Search Terms: Eye Tracking, Landing Performance

Subjects ($n = 16$) performed 15 simulated landing trials and were evaluated on their ability to land on the runway and their eye scanning behavior. Six subjects were classified as experts and 10 as novices. Quality of each landing flown was rated as good or poor, which was determined based on a scoring algorithm. The landing trial started at 1000 ft on a 45 degree turn to final. The approach was approximately 2-3 minutes. Novice subject landings were more varied and were short of the optimal landing point ($p < 0.001$) compared to experts who tended to land past the optimal point. Expert subjects had more fixations ($p = 0.006$), more runway aimpoint ($p = 0.033$), and more airspeed fixations ($p < 0.001$) than novice subjects. Experts had more total fixations than novices ($F[1,191] = 2.816$, [$p = .095$]).

Kim, S., Prinzel, L., Kabar, D. B., Alexander, A. L., Stelzer, E. M., Kaufmann, K., & Veil, T. (2011). Multidimensional measure of display clutter and pilot performance for advanced head-up display. *Aviation, Space, and Environmental Medicine, 82*, 1013-1022.

Search Terms: Adult, Aviation, Computer Graphics, Computer Simulation, Female, Humans, Linear Models, Male, Middle Aged, Psychomotor Performance, User-Computer Interface, Visual Perception, Workload

The study assessed the influence of pilot experience, heads-up display (HUD) configuration/classification, flight segment, and flight workload on pilot perceptions of display clutter, cognitive load, and actual flight performance. Eighteen commercial airline pilots with no prior HUD experience participated in this study. Pilots were asked to fly a standard instrument landing system approach to runway 15R at Reno-Tahoe International airport. The flight included three segments of approach: initial approach fix to glide slope, intercept to final approach fix, final approach fix to decision height. Independent variables included HUD configuration: 3 "high clutter" displays, 3 "medium clutter" displays, and 3 "low clutter" displays. Each pilot flew six trials (three crosswind: high workload and three no wind: low workload).

Dependent variables included pilot rankings of various dimensions of clutter; subjective ratings of overall perceived display clutter; and ratings on the underlying dimensions of clutter. Rankings were used to calculate clutter scores. Pilot also provided ratings of workload using the NASA TLX. Performance was evaluated by collecting altitude RMSE, Speed RMSE, and vertical and horizontal deviations from flight path. There was a significant effect of flight experience on clutter ratings, $F(2, 44) = 3.768$, $p = 0.031$, a significant effect HUD configuration on clutter ratings, $F(2, 44) = 5.043$ $p = 0.011$, and a significant effect of flight workload on clutter ratings, $F(1, 44) = 12.542$, $p = 0.001$. A significant interaction effect between flight experience and HUD configuration on clutter ratings was found, $F(4, 44) = 3.122$, $p = 0.024$. There was a significant effect of HUD configuration on NASA TLX scores, $F(2, 45) = 7.911$, $p = 0.001$. There was a significant effect of flight workload on NASA TLX scores, $F(1, 45) = 10.945$, $p = 0.002$, with higher scores indicating higher flight workload. A significant effect of HUD configuration on altitude RMSE was found, $F(2, 45) = 3.533$, $p = 0.038$. Low clutter configuration was associated with less stable control than medium clutter. There was a significant effect of flight segment on altitude RMSE, $F(1, 45) = 6.254$, $p = 0.014$. Intercept to final approach fix had lower stable control (higher RSME) than the final approach fix to decision height. There was also a significant interaction effect between flight segment and display configuration on altitude RMSE, $F(2, 45) = 8.514$, $p = 0.001$, and a significant effect of HUD configuration on horizontal deviation, $F(2, 45) = 14.51$, $p = 0.001$. Low clutter configuration was associated with less stable control than medium clutter.

Li, W., Horn, A., Sun, Z., Zhang, J., & Braithwaite, G. (2020). Augmented visualization cues on primary flight display facilitating pilot's monitoring performance. *International Journal of Human-Computer Studies, 135,* 102377.

Search Terms: Augmented visualization, Attention Distribution, Flight Deck Design, Human-Computer Interaction, Situation Awareness

Subjects ($n = 20$) completed a flight landing task using both a traditional and a newly developed augmented primary flight display design. The augmented display places a green border around data for the airspeed, altitude, or heading indications during mode changes whereas the traditional display does not. Participants had to call out to an instructor every 10 knot (kt) change in airspeed and every 100 ft change in altitude. A second task required the participants to monitor flight mode annunciation field, wherein they were required to call out any change. Cognitive workload was assessed at the completion of each scenario using the NASA TLX. Eye tracking data were also collected to assess pupil dilation, fixation duration, fixations counts, and mental demand. Situational awareness significantly increased and perceived workload decreased with the augmented flight display, which shortened pilots response time. Pupil dilation ($p < 0.001$), fixation duration ($p < 0.01$), and fixation counts ($p < 0.001$) were different between flight displays. Significant differences between flight displays were observed for NASA TLX mental demand ($p < 0.01$), temporal demand ($p < 0.01$), performance ($p < 0.01$), and effort ($p < 0.01$) but not physical demand ($p = 0.420$) and frustration ($p = 0.179$).

Müller, S., Schreiter, K., Luckner, R., & Manzey, D. (2017). Manual flying and energy awareness. *Aviation Psychology and Applied Human Factors, 7,* 18-27.

Search Terms: Thrust Control, Energy Displays, Energy Awareness, Augmented Flight Control, Pilot Workload

Licensed commercial aircraft pilots ($N = 24$) completed a simulated required navigation performance (RNP) approach pattern under three different visual thrust control display systems in a counterbalanced sequence. The task was presented visually, data were spatial, and response was manual. To increase the demands on percise flight path control, RNP was adjusted from 0.3 to 0.1, steady 15-knot crosswinds were present, and fog obscured the runway until the plane reached 1240 ft. The first condition (conventional) was the control condition where the simulator was configured for manual raw-data flight without any enhanced energy information and no assistance of thrust control. The second flight display condition (energy primary flight display [PFD]) included the conventional concept of thrust control with the energy PFD enriched by the presentation of the total energy angle (TEA) and flight path angle (FPA). The third condition (nxControl) included the energy PFD, internal nxController for thrust and speed brakes control, and the nxStatus display. Performance was measured as deviations from a given reference pattern across time for prescribed altitude, airspeed, and lateral flight path targets. Workload was assessed subjectively using the NASA TLX and objectively with the input activity at the thrust and speed brake lever (reflects how often the pilot has to invest cognitive and physical effort to re-assess and adjust the energy state of the aircraft). Situational awareness was assessed using the Situation Awareness Global Assessment Technique (SAGAT) and Situational Awareness Subjective Workload Dominance (SA-SWORD). The SAGAT was implemented by freezing the flight simulation at one of three possible points of the first approach and blanking the flight displays. Pilots were required to recall the indicated airspeed, lateral RNP deviation, vertical RNP deviation, barometric altitude, vertical speed, pitch, fan rotation speed, and heading. For the SA-SWORD, pilots judged which simulator configuration supported their situational awareness better. Results found that pilots better maintained airspeed with the energy PFD ($P < 0.001$) and nxControl ($P = 0.005$) compared to the conventional condition however a trend towards greater root mean square error for altitude was observed in the nxControl condition compared to conventional and energyPFD ($P = 0.163$). No differences in lateral flight path were observed ($P = 0.147$). Significantly less level activity was observed for the nxControl condition compared to the other conditions ($P < 0.001$) and no difference was present between conventional and energy PFD ($P = 0.299$). No difference in speed brakes were noted between conventional and energy PFD. No diference in total score for the NASA TLX was observed between conditions ($P = 0$; $P = 23$).Physical demand score was significantly lower for energy PFD (36.8) and nxControl (36.2) compared to the conventional condition (45.2; $P = 0.017$). No difference in SAGAT was observed between conditions (P = 0.280). For the SA-SWORD, ratings were significantly higher for the energyPFD (0.61) and nxControl (0.61) than the conventional condition (0.17; $P = 0.016$). No difference was observed between energy PFD and nxControl ($P = 1.00$).

Stanton, N. A., Plant, K. L., Roberts, A. P., Allison, C. K., & Harvey, C. (2018). The virtual landing pad: Facilitating rotary-wing landing operations in degraded visual environments. *Cognition, Technology & Work, 20*, 219-232.

Search Terms: Rotary-Wing, Head-Up Display, Situation Awareness, Workload

Subjects ($N = 13$), who were qualified rotary-wing pilots, completed four experimental flying conditions (clear, clear + HUD, fog, fog + HUD) in which they flew five nautical miles (nm) out to sea and landed on the target helipad. During the fog condition, the pilots were required to complete the landing task in 15 minutes. The HUD provided the following data: conformal compass, heading readout, airspeed indicator, gull wing horizon line, attitude indicator, vertical speed indicator, air speed indicator, wind direction and strength indicator, ground speed, and distance to go. The task was presented visually in a simulator and data were presented visually (traditional heads down display versus HUD) in various weather conditions that degraded vision, and the pilots had to respond manually by landing the helicopter. A familirazation session was completed and a software developer explained each instrument in the HUD and how the symbology worked. A baseline flight in clear weather was performed prior to and following the test conditions. Weather significantly altered the pilots' awareness of desired heading, rate of descent, groundspeed, power status, required landing point, drift, and view of the outside environment. Degraded visual conditions without the HUD reduced awareness of rate of descent groundspeed, power status, required landing point, and outside environment compared to clear conditions. Awareness of rate of descent, drift, and desired heading was higher in clear conditions with the HUD than degraded conditions with the HUD. Pilots' awareness of groundspeed, power status, required landing point, and view of the outside environment was lower in degraded visual conditions with the HUD than in clear conditions with the HUD. Awareness of desired heading and drift was lower in degraded visual conditions without the HUD than in clear conditions without the HUD. Awareness of desired heading, rate of descent, groundspeed, power status, required landing, drift, and outside environment were not affected by display type or interactions between display and weather. Weather significantly effected situtation awareness but not display or the interaction between display and weather. Pilot workload was higher with and without HUD in degraded visual conditions and an even greater effect without the HUD was seen. Pilots were more likely to go-around in degraded visual conditions with and without the HUD than in clear conditions. Mental workload, physical workload, temporal workload, performance, effort, frustration, and overall workload were significantly affected by weather. Lower mental load, physical load, temporal load, effort, and overall workload was observed in clear conditions with the HUD compared to in degraded visual conditions without HUD. In addition, in clear visual conditions with the HUD, lower mental load, physical load, temporal load, and effort were reported compared to similar conditions without the HUD.

Zheng, Y., Lu, Y., Wang, Z., Huang, D., & Fu, S. (2015). Developing a measurement for task complexity in flight. *Aerospace Medicine and Human Performance, 86*, 698-704.

Search Terms: Aerospace Medicine, Aircraft, Computer Simulation, Heart Rate, Humans, Male, Task Performance and Analysis, Workload/Psychology

Subjects (n = 10) completed simulated flight tasks to develop a measurement called Task Complexity in Flight (TCIF). This measurewas implemented across 11 tasks in the takeoff and landing phases of flight, and weree selected from standard operating procedures (SOPs) for the CRJ-200 flight simulator. Actions size complexity, actions logic complexity, information control exchange complexity, and control mode complexity were the components of the TCIF. The Bedford Workload Scale and heart rate were used to measure workload. The results found that as task complexity increased, higher the Bedford Workload Scale ratings were collected (R = 0.851, P = 0.001). TCIF also had a strong correlation with heart rate (R = 0.816, P = 0.002).

This space is intentionally blank.

One Assessment Type—Performance, Physiological, or Subjective.

Babu, M. D., JeevithaShree, D. V., Prabhakar, G., Saluja, K., Paskilkar, A., & Biswas, P. (2019). Estimating pilots' cognitive load From ocular parameters through simulation and in-flight studies. *Journal of Eye Movement Research, 12*(3).

Search Terms: Cognitive Load, Aviation Safety, Eye Gaze Tracking, Fixation, Pupil Dilation, Saccades

Subjects (n = 14) completed a task that involved longitudinal tracking of a target. The boundaries remain in a particular state for 60 seconds and then would shrink by 20% every 60 seconds. The design of the task ensured changes in cognitive load occurred by two horizontal lines constantly closing in together and acting as boundaries. Certain trials required subjects to simulatenously complete a secondary task. The secondary task required the subjects to depress a throttle button in the direction of an arrow that appeared on the display. Three conditions were analyzed. Condition 1 included flying without boundaries or operating the secondary task; condition 2 included flying within the boundaries; condition 3 includedflying within boundaries while performing the secondary task. Eye tracking was assessed during all three conditions. Pupil dilation, saccadic intrusion, and number of saccades and fixations varied significantly with the change in cognitive load. Specifically, pupil dilation in the left [$F(2,36) = 7.18$, $p < 0.05$, $\eta^2 = 0.28$] and right eye [$F(2,36) = 8.67$, $p < 0.05$, $\eta^2 = 0.26$] were significant between all three conditions. Number of saccadic intrusions, [$F(2,36) = 12.28$, $p < 0.05$, $\eta^2 = 0.45$] was significant between conditions 1 and 2 and between conditions 1 and 3. Number of saccades [$F(2,36) = 12.51$, $p < 0.01$, $\eta^2 = 0.63$], were significant between conditions 2, and 3, and conditions 1 and 3. Number of fixations [$F(2,36) = 30.06$, $p < 0.01$, $\eta^2 = 0.49$], were significant between conditions 1 and 2 and conditions 1 and 3.

Çakır, M., Vural, M., Koç, S., & Toktaş, A. (2016). Real-time monitoring of cognitive workload of airline pilots in a flight simulator with fNIR optical brain imaging technology. In *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience* (pp. 147-148). Springer International Publishing.

Search Terms: Mental Workload Estimation, Optical Brain Imaging, fNIR, Neuroergonomics, Linear Discriminant Analysis

Commercial and military pilots (N = 8) with a mean flight experience of 10,712 flight hours (SD = 5057) engaged in four different flight scenarios varying in complexity. During the flight scenarios, the pilots were instrumented with various sensors collecting an assortment of physiological and behavioral metrics for the Advanced Cockpit for the Reduction of Stress and Workload larger study (ACROSS). In this paper, the authors analyzed the functional near infrared spectroscopy (fNIRS) data collected from a workload state classification perspective. No results were reported on the differences between the direct output of the fNIRS system; rather, statistics regarding the classification accuracy of the Linear Discriminant Model were reported and validated. Using this approach, the authors' discriminant analysis method was able to derive two functions that could significantly discriminate workload categories ($p < 0.001$). The first function accounted for 73% of the total variability in the data and discriminates between lower (i.e., 0 level workload task) and higher (i.e., 1 and 2 level workload tasks) workload cases. The second function accounted for 27% of the total variability of the data and discriminates between high workload cases (i.e., 2 level workload task) from lower workload cases (i.e., 0 and 1 level workload tasks).

Dalecki, M., Bock, O., & Guardiera, S. (2010). Simulated flight path control of fighter pilots and novice subjects at +3 Gz in a human centrifuge. *Aviation, Space, and Environmental Medicine, 81*(5), 484-488.

Search Terms: Adult, Aerospace Medicine, Computer Simulation, Germany, Humans, Hypergravity/Adverse Effects, Male, Military Personnel, Motor Skills, Reaction Time, Stroop Test

Subjects ($n = 24$) completed a flight task in a simulator within a centrifuge in which they used a flight control stick to purse a target plane. The path of the target continuously changed altitude while the subjects were exposed to +3 Gz acceleration forces. A modified Stroop task was performed in intervals of 3-7 seconds and subjects were required to report the loudness of each word as quickly as possible. Fourteen subjects were novice pilots and ten were fighter pilots. The experiment consisted of three blocks, with each one having one repetition on the pursuit task (5 minutes), the Stroop task (3 minutes), and a combination of the pursuit and Stroop tasks (5 minutes). The order of the blocks was randomized between subjects and the first block was used as practice, therefore, the data were not analyzed. RMSE between the response and target, oscillation, and mean reaction time were the performance measures that were analyzed. Novice subjects had a significantly greater RMSE than pilots (46% larger in +3 Gz than in +1 Gz) during the first minute. Stroop reaction time was high in novices (+5.27%) than in fighter pilots (+3.77%)($p < 0.05$) and these differences were greater under dual-task conditions. No significant effects were observed for dual task costs (dual task score-single task score/single task score).

Wang, Z., Zheng, L., Lu, Y, & Fu, S. (2016). Physiological indices of pilots' abilities under varying task demands. *Aerospace Medicine and Human Performance, 87*(4), 375-381.

Search Terms: Adult, Aerospace Medicine, Blinking, Heart Rate, Humans, Male, Middle Aged, Professional Competence, Respiratory Rate, Saccades

Commercial airline pilot subjects ($n = 24$) were separated into two groups based on experience, with mean flight hours equaling 11,822 ± 3809.9 for the experienced group and 1,578.8 ± 1059.8 for the less experienced group. Subjects completed simulated flight tasks such as standard instrument departure and standard terminal arrival in good weather, engine failure, and hydraulic systems failure. Subjects were outfitted with a Tobii head-mounted eye tracking system (30 hertz [Hz]) to measure pupil size, pupil position, and gaze position. A Zephyr Technology Bioharness system was used to detect ECG (250 Hz) and respiration (18 Hz). Results showed that more experienced pilots had lower blink rates ($P < 0.01$), heart rate ($P < 0.001$), and respiration rate ($P < 0.01$) than less experienced pilots during both normal tasks and emergency tasks. Fixation duration ($P = 0.036$), saccade rate ($P = 0.65$), and respiration amplitude ($P = 0.69$) were not different between novice and fighter pilots in the normal flight task scenarios. In the emergency flight scenarios, pilot experience significantly affected saccade rate ($P < 0.01$), blink rate ($P < 0.001$), heart rate ($P < 0.01$) and respiration rate ($P < 0.01$).

Again, fixation duration ($P = 0.07$) and respiration amplitude ($P = 0.33$) were not different between experienced and less experienced pilots in emergency situations. Interestingly, task demand did not show significant overall effect in either more experienced ($P = 0.14$) nor less experienced ($P = 0.60$) pilots. A principal component analysis revealed two factors that accounted for 73.02% of the total variance, with the first factor (38.50% of variance; representing pilot physical fitness) loaded heavily on by heart rate and respiration rate and the second factor (34.52% of variance; representing pilot information access strategy) loaded heavily on by saccade rate and blink rate. This study stands as a reference for the effect of experience on physiological metrics of cognitive workload and identifying the factors that contribute to the observed changes in physiology while under cognitive stress.

Zhang, X., Xue, H., Qu, X., & Li, T. (2017). Can fixation frequency be used to assess pilots' mental workload during taxiing? In *Lecture Notes in Computer Science* (pp. 76-84). Springer International Publishing.

Search Terms: Mental Workload, Pilot, Fixation frequency, Fixation, Eye Movement

Subjects ($n = 3$) were asked to land from the last approach point and sucessfully taxi to the gate. Visibility varied from a clear day, night, and low visibility with two different difficulty taxiing ways (simple versus difficult). The simple taxiing was shorter in length and had less turns than the difficult taxiing condition. The number of fixations were assessed with eye tracking. A significant effect was found between scenarios ($F(5,17) = 5.365$, $P < 0.05$). No difference was observed between the level of difficulty of taxiing or visability.

Zheng, Y., Lu, Y., Yuwen, J., & Shan, F. (2019). Predicting workload experienced in a flight test by measuring workload in a flight simulator. *Aerospace Medicine and Human Performance, 90*(7), 618-623.

Search Terms: Adult, Aerospace Medicine/Methods, Aircraft, Decision Making, Humans, Male, Middle Aged, Pilots/Psychology, Simulation Training, Workload/Psychology

Subjects ($n = 13$) completed both a simulator flight test and a live flight test in an ARJ21-700 aircraft. Three flight scenarois were performed including standard instrument departure (SID), standard instrument approach (SIA), and direct mode approach (DMA). Subjects wore a head-mounted eye tracker to assess blink rate, and heart rate and respiration rate were measured with a Bio Harness. The Bedford Workload Scale was completed during each flight scenario and the NASA TLX was completed after each flight. Both simulator and live flight scenarios [$F(2,24) = 31.968$, $P < 0.001$] and flight environment [$F(1,36) = 9.842$, $P = 0.002$] had a significant influence on workload, but the interaction of these two factors was insignificant [$F(2,24) = 0.074$, $P = 0.929$]. In the SID scenario, the correlation of workload in the flight simulator and during the flight test was moderate ($R = 0.529$, $P = 0.077$). However, the greatest correlations were found in the SIA scenario ($R = 0.808$, $P = 0.001$) and DMA scenario ($R = 0.815$, $P = 0.001$), and the overall correlation was also significant ($R = 0.864$, $P < 0.001$). For the Bedford Workload Scale, neither flight scenario [$F(2,24) = 0.619$, $P = 0.542$] nor flight environment [$F(1,36) = 0.048$, $P = 0.828$] exhibited significant influence on workload. The overall correlation of workload in the flight simulator and during the flight test was weak ($R = -0.151$, $P = 0.379$).

Intermediate Population

All Assessment Types—Performance, Physiological, and Subjective.

Diaz-Piedra, C., Rieiro, H., Cherino, A., Fuentes, L., Catena, A., & Di Stasi, L. (2019). The effects of flight complexity on gaze entropy: An experimental study with fighter pilots. *Applied Ergonomics, 77,* 92-99.

Search Terms: Brain Activity, Cognition, EEG, Eye Movements, Eye Tracking, Flight

Subjects (n = 14) completed a total of eight simulated flight tasks of varying levels of difficulty (low, medium, and high). The two longest flights were low complexity recognition flights without the presentation of an emergency situtation. The three medium complexity flights included multifunction display failure simultaneously with metal fragments in the engine oil, flying with on engine inoperative circuit, and with one engine on fire. The three high complexity flights required pilots to operate an auto-rotation landing, flying simultaneously with one inoperative engine circuit and suffering a hydraulic failure while performing a crosswind landing, and landing while suffering a tail rotor failure. Gaze entropy was statistically lower in the emergency flight exercises than low complexity ($p < 0.05$). The power spectra in theta band at the frontal and the central channels were lower during the low complexity flights compared to the emergency flights (frontal channels: 3.36 ± 0.54 vs 3.79 ± 0.83 and 3.78 ± 0.87; central channels: 1.60 ± 0.29 vs. 1.74 ± 0.32 and 1.71 ± 0.32, corrected p-values < .05). Significantly higher perceived task workload was reported for performing the emergency flights (for medium: 38.57 ± 10.70 and high complexity flights: 41.16 ± 10.63) compared to the recognition flights (low complexity flights: 21.77 ± 10.24); corrected p-values < .05. Pilot's performance during the low complexity flights was significantly higher (100% ± 0.0 of required maneuvers performed) compared to the emergency flights (medium and high complexity flights pilots performed 87.99% ± 6.87 and 84.18% ± 7.23, respectively).

Goncalves-Martins, A. P., Hölscher, S., & Dautermann, T. (2020). Evaluation of a tunnel-in-the-sky head-up display design for curved approaches using eye-tracking. *Aviation Psychology and Applied Human Factors, 10*(1), 3-12.

Search Terms: HUD, Eye-Tracking, Tunnel-in-Sky, Curved Approach

Subjects held professional pilots liscense with an instrument rating (n = 6; M = 29, SD 4.38) with flight experience ranging from 148 to 2,800 hrs. Two subjects had HUD experience; the subject pool experience ranged from 0 to 100 hr (M = 17.8, SD 40.3). This experiment was conducted in a fixed-base generic cockpit simulator with Airbus cockpit hardware including side sticks, tillers, thrust levers, flap and gear handles, rudder and brake pedals. Two different display layouts were tested: "tunnel-in-the-sky" and standard HUD. The "tunnel-in-the-sky" display presents path guidance within a tunnel with a centerline; pilots are expected to maintain position along this centerline. If the pilot maintains this course, flight information is minimized to the side that moves sychronisly with the pilot. The standard display features path guidance provided solely by deviaion bars with dial guages on either side.

Pilots flew approaches to runway 26 at the research airport. Pilots flew once with the "tunnel-in-the-sky" layout and once with the standard layout, counterbalanced. Dependent variables included proportion of fixations (number of fixations on an area of interest [AOI] divided by total number of fixations), average dwell time, lateral deviations from centerline, departures from vertical descent profile, NASA TLX, and a 3-dimensional situational awareness rating technique (3D-SART). Wilcoxon signed-ranked tests were used for pairwise comparison of displays. Overall NASA-TLX scores showed a significant decline in workload scores with the "tunnel in the sky" layout vs the standard display layout ($Z = 2.20$, $p = 0.028$, $r = 0.64$). All other scales within the NASA TLX with the exception of perormance and frustration were significanltuy higher with the standard display ($p < .05$). The 3D-SART found a signinifcantly higher demand for "Supply of Attentional Resources" ($Z = 2.06$, $p = 0.039$, $r = 0.59$) and "Demand for Attentional Resources" ($Z = 2.04$, $p = 0.041$, $r = 0.59$) with the standard display VS "tunnel in the sky" layout. Performance data found that there were significantly less verticle deviations found with the "tunnel-in-the-sky" layout ($M = 3.2$, SD 1.34) VS the standard display ($M = 9.81$, SD 4.31) $Z = 2.02$, $p = 0.028$. Lateral deviations from the flight path were statistically lower with the "tunnel-in-the-sky" layout ($M = 9.2$, SD 0.91) VS the standard display ($M = 60.42$, SD 20.07) $Z = 2.02$, $p = 0.043$. Eye tracking data showed a higher proportion of fixations in the bore area with the "tunnel in the sky" display design ($Z = 2.20$, $p = 0.043$) as well as longer dwell time in the bore area ($Z = 2.02$, $p = 0.043$).

Mansikka, H., Virtanen, K., & Harris, D. (2019). Comparison of NASA-TLX scale, Modified Cooper-Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks. *Ergonomics, 62*(2), 246-254.

Search Terms: Adult, Aviation/Statistics and Numerical data, Computer Simulation, Female, Heart Rate, Humans, Male, Pilots/Psychology, Task Performance and Analysis, United States, United States National Aeronautics and Space Administration, Work/physiology, Workload/Psychology, NASA TLX Scale, Pilot Mental Workload, Inter-Beat Interval, Modified Cooper-Harper Scale

Finnish Air Force McDonnell-Douglas F/A-18C pilots ($N = 27$) engaged with a weapon tactics and situational awareness trainer (WTSAT; a flight simulator used for basic and advanced flight training) to conduct eight instrument landing system (ILS) approaches with varying temporal demand levels. Temporal demand was manipulated by varying the starting ranges from 15 NM (27.8 km) to 8 NM (14.8 km) in eight 1 NM (1.9 km) intervals, reducing the overall time available from 6 minutes 22 seconds to 3 minutes 27 seconds with 25.5 second steps. Additionally, the ILS approaches were each complemented with 20 additional tasks, regardless of the duration of the task. During task engagement, ECG data was recorded using a Mind Media NeXus-10 MKII system and the mean interbeat interval (IBI) was calculated using the raw ECG signal. Subjective measures were recorded after each trial using the Modified Cooper-Harper (MCH) and the NASA TLX. Results showed that the cognitive workload assessments utilized showed significant differences across the three performance categories defined by high, medium, and low performance on the ILS task. The IBI, MCH, and NASA TLX were able to differentiate the high-performance categories from the medium- and low-performance categories; while the MCH and NASA TLX were also able to differentiate between medium- and low-performance categories. A strong, positive correlation between MCH and NASA TLX reveal that each scale is providing a similar result; however, the lack of correlation between IBI values and MCH and NASA TLX scores does reveal workload metric dissociation.

Yu, C., Wang, E., Li, W., & Braithwaite, G. (2014). Pilots' visual scan patterns and situation awareness in flight operations. *Aviation, Space, and Environmental Medicine, 85*(7), 708-714.

Search Terms: Adult, Aerospace Medicine, Attention/Physiology, Aviation, Awareness/Physiology, Computer Simulation, Eye Movements/Physiology, Fixation, Ocular/Physiology, Humans, Male, Military Personnel, Pupil/Physiology, Taiwan, Task Performance and Analysis, Workload

Subjects (n = 18) were military fighter pilots and they performed flight maneuvers of varying levels of difficulty in a simulator. In addition to operating the aircraft they had to follow the navigation system and enter code by using varying flight deck interfaces. Subjects had to simultaneously intercept the proper route and turn toward the target. A flight instructor evaluated each pilot's performance and also recorded their situational awareness by activating the generator malfunction light during the highest workload phase. The pilot was instructed to respond to this stimulus by pushing the master caution light button and call "generator out." Eye tracking was used to assess percentage of fixation, duration of fixation, and pupil diameter. Subjective workload was also assessed. Pilots distributed 59.92% of their fixations on the HUD (AOI-1) and 39.18% (arc sine value = 38.75) outside of the cockpit (AOI-5). Fixation durations on the HUD and outside of the cockpit were greater than on AOI-2: the integrated control panel (ICP), AOI-3: Right Multiple Function Display (RMFD), and AOI-4: Left Multiple Function Display (LMFD). Pupil size at the phase of aiming was the largest, followed by release and break-away, and then preparation. Pilots who identified the warning light had better situational performance and significantly lower workload.

This space is intentionally blank.

Combination of Two Assessment Types—Performance, Physiological, and Subjective.

Hancock, P. (2013). Task partitioning effects in semi-automated human-machine system performance. Ergonomics, 56(9), 1387-1399.

Search Terms: Aerospace Medicine, Aircraft/instrumentation, Automation, Computer Simulation, Ergonomics, Female, Humans, Male, Man-Machine Systems, Reaction Time, Task Performance and Analysis

Professional pilots ($n = 12$) performed a flight simulation consisting of three component sub-tasks: (i) tracking, (ii) monitoring, and (iii) targeting. Each participant completed a simulation five times: three in fully manual condition, one in partial automation condition, and one in fully automated condition. Experimental design was a completely within-subjects. Prior to the flight performance, participants completed the Subjective Workload Assessment Technique (SWAT) card sort procedure (Reid and Nygren, 1988) and then the experiment began. At the end of each trial, the pilot was asked for their concluding SWAT response. For tracking, dependent measures included RMS error, response time, percentage of responses made, and percentage of those responses that were correct. For targeting, dependent measures included percentage of responses made, percentage of those responses that were correct, and mean time for response. Missed targets and false alarms were also recorded. There was a significant difference between manual and automated conditions ($p < .01$) in RMS error. There was a significant difference between partial and fully automated conditions ($p < .05$). The fully automated condition overall resulted in lower RMS error and fully manual had the highest RMS error. There was a significant difference in the mean time for response during targeting between the partial automation and fully automated conditions ($p < .05$). Response time less during partial automation. There were no significant differences in any measure of the SWAT.

Hsu, C., Lin, S., & Li, W. (2015). Visual movement and mental-workload for pilot performance assessment. In *Engineering Psychology and Cognitive Ergonomics* (pp. 356-364). Springer International Publishing.

Search Terms: Eye Movement, Pilot Performance, Flight Safety, Attention Allocation

Subjects ($n = 18$) completed an air-to-surface training course where an instructor rated their performance and eye tracking data was assessed. In a simulator, subjects were given scenarios that required them to dive in at high speed to the expected bomb release point or firing altitude for the key operating stages from flatten, tracking, pick-off, to off. High performing pilots had significantly greater total time in zone ($t = -2.83$, $p < .05$), percent time in zone ($t = -3.18$, $p < .001$), fixation counts ($t = -2.49$, $p < .05$), percentage of total fixations ($t = -2.27$, $p < .05$), total fixation duration ($t = -2.51$, $p < .05$), and gaze-point counts ($t = -2.48$, $p < .05$) compared to low performing pilots. A significant interaction between flight performance and experience with fixation duration was found ($F = 4.89$, $p < .05$). Subjects demonstrating high performance exhibited a significantly shorter glance distance than low performing subjects ($t = 2.46$, $p < 0.05$). Pupil size at the flatten and pick-off stages were significantly larger than that at the off stage.

Mansikka, H., Virtanen, K., & Harris, D. (2019). Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Transactions on Human-Machine Systems, 49,* 1-9.

Search Terms: Task analysis, Aircraft, Training, Heart Rate, Atmospheric Modeling, Military Aircraft, Instruments

Subjects (n = 37) completed two approach and two commits in a simulated air combat mission. In the high demand approach (Approach-HI) subjects flew to a planned landing airfield after the mission's combat phase while following the radar track of a leading aircraft and maintaining a predefined separation distance. The lead aircraft reduced its speed at the last possible moment to limit the time that the subjects could prepare for the approach. Additionally, subjects had to respond to tactical radio calls. In the low demand approach (Approach-LO), the subjects were allowed to fly without a lead aircraft, allowing them to control their airspeed and flight profile. Significant differences in performance were observed between the Approach-HI (M = 69.1, SD = 17.4) and Approach-LO (M = 82.1, SD = 8.3); $t(31)$= -3.886, $p < 0.001$. A significant difference in performance was also found between Commit-LO (M = 57.0, SD = 10.4) and Commit-HI (M = 44.8, SD = 8.0); $t(31)$ = 5.561, $p < 0.01$. There was also no overall significant difference in HR values between HI and LO segments (F1,31 = 2.275, p = 0.142). Mean HR values for LO segments was 95.1; mean HR values for HI segments was = 94.0. No difference in workload (as as sessed by IBI values) between the main effects of Approach/Commit or HI/LO segments. IBI values in Approach-HI were significantly lower than in Approach-LO, therefore Approach-HI had a higher task demand than Approach-LO.

Mansikka, H., Virtanen, K., Harris, D., & Simola, P. (2016). Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test. *Applied Ergonomics, 56,* 213-219.

Search Terms: Pilot Mental Workload, Heart Rate, Pilot Performance

Subjects (n = 26) completed a check ride flight, in a simulator, that consisted of seven segments. The following segments were performed and performance was assessed by a single, expert examiner: 'Takeoff and Ingress,' 'Maneuvering,' 'Level Turns,' 'Single Engine Maneuvering (SEM),' 'VOR (VHF Omni Directional Radio Range) Approach,' Instrument Landing System (ILS) Approach,' and 'Precision Approach Radar (PAR) Approach.' Heart rate and heart rate variability data were collected during flight to assess workload during each segment. The highest task load was expected during instrument approaches without the use of autopilot and the 'Maneuvering' segment was expected to have the lowest task load. Subjects performance was scored on the ILS approach based on deviations from the target airspeed, glide slope, and localizer. Whereas the VOR approaches were scored based on deviations from the target speed, step down fixes, and the final approach course. Performance remained stable between the different mission segments but changes in HR and in the components of HRV were observed. HR and the HRV components were able to differentiate six mission segment pairs that did not have performance differences.

One Assessment Type—Performance, Physiological, or Subjective

Bezerra, F., & Ribeiro, S. (2012). Preliminary study of the pilot's workload during emergency procedures in helicopters air operations. *Work, 41*(1), 225-231.

Search Terms: Adult, Aircraft, Emergencies/Psychology, Humans, Man-machine Systems, Mental Processes/Physiology, Middle Aged, Qualitative Research, Task Performance and Analysis, Time Factors, Workload/Psychology

Subjects ($n = 10$) completed live flight training exercises aimed at manually performing corrective actions to four different emergency situations that were triggered by a flight instructor. The tasks were as follows: Task 1: Emergency response procedures for approach and landing without the operation of the hydraulic system, Task 2: Emergency procedure for approach and landing with tail rotor control failure, Task 3: Emergency procedure for approach and landing with engine failure, and Task 4: Emergency procedure for approach and landing with complete loss of tail rotor effectiveness. Subjects completed the NASA TLX after performing each task; however, only mental demand score and physical demand score were used in the final analysis. Mental demand was highest for tasks two, three, and four, whereas task one had the highest physical demand.

Di Stasi, L., Diaz-Piedra, C., Suarez, J., McCamy, M., Martinez-Conde, S., Boca-Dorda, J., & Catena, A. (2015). Task complexity modulates pilot electroencephalographic activity during real flights. *Psychophysiology, 52*(7), 951-956.

Search Terms: Adult, Aircraft, Aviation, Brain Waves, Electroencephalography, Humans, Male, Military Personnel, Task Performance and Analysis,Young Adult, Descriptors, EEG, Fatigue, Neuroergonomics, Safety, Simulation, Training

Pilots ($N = 8$) with an average of 160 flight hours ($SD = 25$) engaged in a flight task using a Sikorski S-76C helicopter while instrumented with a portable EEG recorder (SOMNOwatch plus EEG-6; 256 Hz). The flight lasted about 60 minutes for each subject and consisted of four 15-minute stages that corresponded with the main phases of flight including take off procedures, two airwork procedures, and landing procedures. Results showed a significant interaction between flight stage and frequency band ($p < 0.01$). With this interaction, the authors observed that high-frequency EEG bands (alpha and beta) were able to discriminate between flight stages, while low-frequency bands were unable to differentiate between flight stages. This study demonstrates the ability to use a mobile EEG platform to collect data in a real-flight environment with the ability to use changes in high-frequency EEG data as a method to identify changes in task demand.

This space is intentionally blank.

Mansikka, H., Simola, P., Virtanen, K., Harris, D., & Oksama, L. (2016). Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. *Ergonomics, 59*(10), 1344-1352.

Search Terms: Aerospace Medicine/Methods, Aircraft, Analysis of Variance, Aviation, Computer Simulation, Finland, Heart Rate/Physiology, Humans, Military Personnel, Pilots, Task Performance And Analysis, Workload, Heart Rate, Heart Rate Variation, Performance, Pilot Mental Workload

Subjects (n = 35) completed 12 full test procedures that all included and instrument landing system approach with varying levels of task demand. Task demand was manipulated by varying the starting times of the trials. Subtasks such as reading back the clearance, checking requested flight parameters and reporting to air traffic control, setting the cockpit instrument to a directed value, initiating a related emergency procedure, selecting an anti-ice switch to on, resetting a fuel level warning, calculating landing speed based on indicated fuel state, switching to an alternative display, tuning a radio preset to an indicated frequency, copying a value to a knee pad, making a full stop landing, as well as maintaining an approach course, guidepath and minimum approach speed. Heart rate, heart rate variability, and performance were examined. Each trial was rated as having high performance, sub-standard performance, or low performance. There were statistically significant overall HR/HRV differences between performance categories; $F(7,21)=3.9$, p<0.05, n^2, p=0.94. Significant HR/HRV differences between performance categories were found on: mean respiration rate; $F(3,81) = 47.1$, $p < 0.001$, n^2 $p = 0.64$; Standard deviation (*SD*) of successive normal QRS complexes (normal-to-normal [NN]) $F(3,81) = 6.5$, $p < 0.01$, n^2 $p = 0.19$; mean heart rate; $F(3,81) = 31.6$, $p < 0.01$, n^2 $p = 0.54$; number of successive NN interval pairs that differ more than 50 ms (NN50) $F(3,81) = 18.1$, $p < 0.001$, n^2 $p = 0.40$; total number number of all NN intervals (pNN50) $F(3,81) = 8.4$, $p < 0.01$, n^2 p =0.24; trianuglar index (HRVTRI) $F(3,81) = 17.2$, $p < 0.001$,n^2, $p = 0.38$. All HR/HRV components showed significant difference between the baseline rest and the high performance category ($p < 0.05$). mean respiration rate ($p < 0.01$), and mean heart rate ($p < 0.01$) were able to differentiate the high performance from the sub-standard performance category. HR/HRV components did not differentiate between the sub-standard and the low performance category.

This space is intentionally blank.

Novice Population

All Assessment Types—Performance, Physiological, and Subjective

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports, 7*(1), 5222.

Search Terms: Adult, Brain Mapping/Methods, Cerebrovascular Circulation, Executive Function/Physiology, Female, Functional Neuroimaging/Methods, Hemoglobins/Metabolism, Humans, Male, Oxyhemoglobins/Metabolism, Pilots, Prefrontal Cortex/Diagnostic Imaging/Physiology, Spectroscopy, Near-Infrared/Methods, Task Performance And Analysis, Workload, Young Adult

Subjects ($n = 26$) engaged in a flight simulator landing task across two levels of difficulty (easy and difficult) and mentally demanding laboratory tasks (executive function tasks outside the aviation domain, not discussed here) while having their blood oxygenation levels monitored using an fNIRS system. The flight task consisted of landing while using an instrumented landing system in an easy condition with perfect visibility and no crosswind and a difficult condition with no external visibility with a strong crosswind. While participants engaged in each task scenario, the hemodynamics (HbO2 signal) of their prefrontal cortex was monitored using a Biopac fNIR100 system and their task performance was captured by the simulator. Following each landing scenario, subjects completed a subjective workload evaluation using a 1-7 scale (undefined scale). Results showed a significant difference between easy and difficult landing scenarios in performance data with trajectory deviations being higher during the difficult condition ($F[1, 25] = 12.69$, $p < 0.01$) and subjective data (F[1, 25] = 73.72, $p < 0.001$). The fNIRS-derived HbO2 signal revealed no significant effect for difficulty ($p > 0.05$); however, results indicate that the period of time during the landing (early and late phase) significantly affected HbO2 signals ($F[1, 25] = 115.85$, $p < 0.001$). Higher HbO2 concentration was present during the late phase of the landing. Additionally, a main effect of optode placement ($F[15, 375] = 9.19$, $p < 0.001$) revealed areas of increased activity (such as in the dorsolateral prefrontal cortex). Continued analysis was conducted to correlate laboratory task performance with piloting performance, but no strong correlation was found across conditions ($r < 0.40$). This study demonstrates the ability of fNIRS to capture variations in cognitive effort in an applied domain, and cautions that cognitive effort as measured by fNIRS alone may be unrealistic.

Kallus, K., Tropper, K., & Boucsein, W. (2011). The importance of motion cues in spatial disorientation training for VFR-pilots. *The International Journal of Aviation Psychology, 21*, 135-152.

Search Terms: NASA TLX SCALE, ECG, EOG, EEG, BSKE Mood Scaling, Aviation, Illusions

The study investigated the role of motion cues for spatial disorientation training in general aviation. Subjects ($n = 42$) included visual flight rules (VFR) pilots randomly assigned to one of three conditions: training-motion ($n = 15$), training no-motion ($n = 15$), and control-motion ($n = 12$). Subjects flew five different profiles: take-off with pitch-up illusion, spin recovery, unusual attitude recovery (UAR), VFR at minimal weather condition, changing runway width and slope. Dependent variables for performance included time to regain safe flight parameters for UARs, instructor rating of pilot performance, self-rating of performance, and NASA TLX rating of performance. Psychophysiological dependent variables included NASA TLX scores, Questionnaire on psychic and physical state (using the "Befindlichkeitsskala"—a standardized German mood scale [BSKE] and the "mehrdimensionale körperliche Symptomliste" [MKSL] translated as the Multidimensional Bodily Symtpom Checklist), ECG (HR and HRV), EOG, EEG. Results indicated there were positive training effects for takeoff with pitch-up illusions, $F(2, 39) = 6.68$, $p = 0.003$. There were positive training effects for inadvertent flight into instrument meterological conditions (IMC), $F(2, 39) = 5.14$, $p = 0.010$. Positive training effects for spin recovery, $F(2, 39) = 4.87$, $p = 0.013$. A significant interaction effect between flight profile and training group was found, $F(8, 156) = 2.51$, $p = 0.014$. Training group had asignificant effect on instructor ratings, $F(2, 39) = 8.47$, p = 0.001. Training group had a significant effect on NASA TLX Performance, $F(2, 39)$, $p = 0.009$. Motion based training better than no-motion training. Training group had significant effect on mental workload, $F(2,39)$, $p = 0.038$. Results indicated the superiority of motion-based over fixed-base simulators for training to cope with illusions resulting from the combination of visual and vestibular cues.

This space is intentionally blank.

Combination of Two Assessment Types—Performance, Physiological, and Subjective.

Hidalgo-Muñoz, A., Mouratille, D., Matton, N., Causse, M., Rouillard, Y., & El-Yagoubi, R. (2018). Cardiovascular correlates of emotional state, cognitive workload and time-on-task effect during a realistic flight simulation. *International Journal of Psychophysiology, 128*, 62-69.

Search Terms: Adult, Aircraft, Computer Simulation, Electrocardiography, Emotions/physiology, Executive Function/physiology, Heart Rate/physiology, Humans, Male, Psychomotor Performance/physiology, Simulation Training, Stress, Psychological/physiopathology, Young Adult, Cognitive Workload, Emotion, Flight Simulation, Heart Rate, Heart Rate Variability, Neuro-Ergonomics

Subjects ($n = 20$) completed two dual-task scenarios while completing a pre-established filght plan in a simulator and a secondary task based on target stimulus discrimination with either a low or high cognitive workload. Additionally, the dual-task scenarios were carried out with or without social stressors. For the low cognitive workload condition, the subjects pressed a screen if the numbers they heard met a simple rule (in the first dual task the numbers were greater than five, in the second dual task the numbers were even). A series of random numbers were displayed on the screen but subjects were instructed to ignore them. For high cognitive workload, the attribute values depended on the color of the visualized number and the previously mentioned rules. In the low arousal condition the subjects were left alone in the flight simulator. In the second dual task condition, considered as the high arousal condition, the participants were filmed and their voice was recorded. Subjects were also monitored closely by two researchers and were told they were in a competition against other subjects. The results indicated cognition or emotion, or the interaction between the factors, had an effect on flight performance. Significant main effects for both emotion and cognition factors were found, with faster reactions times (RT) (less time to respond) for the high arousal condition than for the low arousal condition, $F(1,10) = 13.64$, $p = 0.004$, $\eta p2 = 0.58$, and for low arousal compared to high arousal condition, $F(1,10) = 66.02$, $p < 0.001$, $\eta p2 = 0.87$. No significant effect of emotion or cognition was found on global errors. A main effect of cognition was found for HR, $F(1,19) = 4.56$, $p = 0.046$, $\eta p2 = 0.19$, showing a greater value for high cognitive workload ($M = 86.55$ bpm, $SD = 15.18$) compared to the low cognitive workload ($M = 85.14$ bpm, $SD = 15.47$) condition ($p = 0.013$). Neither cognition or emotion effected HRV. An interaction between the time-on-task and cognition was found in HR, $F(2,38) = 3.39$, $p = 0.04$, $\eta p2 = 0.15$, proportion of R-R intervals which differ more than 20 (pNN-20), $F(2,38) = 6.11$, $p = 0.01$, $\eta p2 = 0.24$, pNN-50, $F(2,38) = 5.66$, $p = 0.01$, $\eta p2 = 0.23$ and pNN-SD, $F(2,38) = 7.71$, $p = 0.001$, $\eta p2 = 0.29$.

This space is intentionally blank.

Scannella, S., Peysakhovich, V., Ehrig, F., Lepron, E., & Dehais, F. (2018). Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft. *Human Factors, 60*(7), 922-935.

Search Terms: Adult, Aircraft, Electrocardiography, Eye Movement Measurements, Female, Humans, Male, Man-machine Systems, Pilots, Psychomotor Performance/Physiology, Young Adult, ECG, Aircraft Pilots, Classification, Eye-tracker, Workload

Participants (n = 11) completed a simulated flight that consisted of three phases. Phases consisted of takeoff (period of 60 seconds, starting from power setting or touch-and-go), downwind (period of 60 seconds, in the middle of downwind), and landing (period of 60 seconds, before touch down). NASA TLX scores were recorded after each phase to determine workload. The results indicated that there were significant differences in workload between the three flight phases (F[2, 20] = 65.3; $p < .001$). Participants reported the highest workload during the landing phase (57.1 ± 5.6) followed by the takeoff (45.4 ± 6.1). Heart rate and heart rate variability were impacted by flight phase ($p < .01$ and $p < .05$, respectively). Lower heart rate was observed for downwind compared to the two other phases ($p < .01$ for both comparisons). No differences in heart rate were observed between the landing and takeoff phases. Saccadic rate and visual entropy were also affected by the flight phase ($p < .001$, for both tests).

This space is intentionally blank.

One Assessment Type—Performance, Physiological, or Subjective.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., & Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PLoS One, 10*(3), e0121279.

Search Terms: Adult, Aerospace Medicine, Aircraft, Aviation, Brain-computer Interfaces, Computer Simulation, Female, Hemodynamics, Humans, Male, Memory, Short-Term/Physiology, Prefrontal Cortex/Physiology, Spectroscopy, Near-infrared, Workload

Visual flight rules (VFR) pilots (N = 19) engaged in an auditory working memory task inside a flight simulator while instrumented with a Biopac functional near-infrared spectrometer module (fNIR100). The task within the simulator involved pilots working with air traffic control (ATC) messages across two demand levels to obtain data specifying values to be input into the autoflight system using four knobs (speed, heading, altitude, and vertical speed). In the low workload condition, the subjects were presented with a similar set of data using the same major digit across all four values. In the high workload condition, subjects were presented more variable/complex inputs for the task, increasing the demand on working memory. Results indicate that the fNIRS system was able to find a significantly higher peak response ($p < 0.001$) in HbO_2 in the high workload condition. Additionally, analysis revealed that the load effect was not homogenous across the measured area ($p < 0.001$), and a localized maximum load effect was present in the right dorsolateral prefrontal cortex. The authors utilized the collected data to estimate the cognitive state (on-task or not-on-task) of the subject using a Moving Average Convergence Divergence (MACD) filter and support vector machine based classification algorithms, demonstrating significantly better than chance classification results.

Han, S., Kim, J., & Lee, S. (2019). Recognition of pilot's cognitive states based on combination of physiological signals. *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, 1-5.

Search Terms: Aviation, EEG, ECG, EDA, PPMs, HCI

Subjects (N = 20) with over 100 hours of flight experience engaged in a Cessna 172 flight simulator task across three paradigms used to invoke distinct mental states: mental fatigue, distraction, and mental workload. The paradigm of interest was the mental workload paradigm defined by presenting subjects three increasingly demanding levels of flight instruction complexity. Data was collected across various physiological measures, including EEG (250 Hz; using the international 10-20 system), ECG, electrodermal activity (EDA), and respiration. EEG results showed a significant decline across all frequencies in each area except in the pre-frontal area. Reported results in other measures indicate an increase in the SDNN (standard deviation of NN intervals) ECG feature, an slight decrease in RF1 and RF2 respiration features, and a large decrease in the mean EDA and EDAsymp features with higher workload state, but none were reported as statistically significant. These results were used in conjunction to classify the pilots' cognitive states, with generalization applicable to the mental fatigue and mental workload states. Using multiple measures in conjunction with canonical correlation analysis yielded higher classification performance than concatenated multi-modality feature analysis.

References

Babu, M. D., JeevithaShree, D. V., Prabhakar, G., Saluja, K., Paskilkar, A., & Biswas, P. (2019). Estimating pilots' cognitive load From ocular parameters through simulation and in-flight studies. *Journal of Eye Movement Research, 12*(3).

Bezerra, F., & Ribeiro, S. (2012). Preliminary study of the pilot's workload during emergency procedures in helicopters air operations. *Work, 41*(1), 225-231.

Cain, B. (2007). *A Review of the Mental Workload Literature*. Defence Research and Development Canada Toronto.

Çakır, M., Vural, M., Koç, S., & Toktaş, A. (2016). Real-time monitoring of cognitive workload of airline pilots in a flight simulator with fNIR optical brain imaging technology. In *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience* (pp. 147-148). Springer International Publishing.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports, 7*(1), 5222.

Cheung, B., Craig, G., Steels, B., Sceviour, R., Cosman, V., Jennings, S., & Holst, P. (2015). In-flight study of helmet-mounted symbology system concepts in degraded visual environments. *Aerospace Medicine and Human Performance*, *86*(8), 714-722.

Cheung, B., McKinley, R. A., Steels, B., Sceviour, R., Cosman, V., & Holst, P. (2015). Simulator study of helmet-mounted symbology system concepts in degraded visual environments. *Aerospace Medicine and Human Performance, 86*(7), 588-598.

Dahlstrom, N., Nahlinder, S., Wilson, G. F., & Svensson, E. (2011). Recording of psychophysiological data during aerobatic training. *The International Journal of Aviation Psychology, 21*, 105-122.

Dalecki, M., Bock, O., & Guardiera, S. (2010). Simulated flight path control of fighter pilots and novice subjects at +3 Gz in a human centrifuge. *Aviation, Space, and Environmental Medicine, 81*(5), 484-488.

De Waard, D. (1996). *The Measurement of Drivers' Mental Workload* [Unpublished Doctoral Dissertation]. University of Groningen.

Di Stasi, L., Diaz-Piedra, C., Suarez, J., McCamy, M., Martinez-Conde, S., Boca-Dorda, J., & Catena, A. (2015). Task complexity modulates pilot electroencephalographic activity during real flights. *Psychophysiology, 52*(7), 951-956.

Diaz-Piedra, C., Rieiro, H., Cherino, A., Fuentes, L., Catena, A., & Di Stasi, L. (2019). The effects of flight complexity on gaze entropy: An experimental study with fighter pilots. *Applied Ergonomics, 77*, 92-99.

Diaz-Piedra, C., Rieiro, H., Suarez, J., Rios-Tejada, F., Catena, A., & Di Stasi, L. (2016). Fatigue in the military: Towards a fatigue detection test based on the saccadic velocity. *Physiological Measures, 37*(9), N62-N75.

Eichinger, A. & Kellerer, J. (2014). Between laboratory and simulator: A cognitive approach to evaluating cockpit interfaces by manipulating informatory context. *Cognition, Technology & Work, 16*, 417-427.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., & Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PLoS One, 10*(3), e0121279.

Goncalves Martins, A. P., et al. (2020). Evaluation of a tunnel-in-the-sky head-up display design for curved approaches using eye-tracking. *Aviation Psychology and Applied Human Factors, 10*(1), 3-12.

Han, S., Kim, J., & Lee, S. (2019). Recognition of pilot's cognitive states based on combination of physiological signals. *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, 1-5.

Hancock, P. (2013). Task partitioning effects in semi-automated human-machine system performance. *Ergonomics, 56*(9), 1387-1399.

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

Hidalgo-Muñoz, A., Mouratille, D., Matton, N., Causse, M., Rouillard, Y., & El-Yagoubi, R. (2018). Cardiovascular correlates of emotional state, cognitive workload and time-on-task effect during a realistic flight simulation. *International Journal of Psychophysiology, 128*, 62-69.

Hsu, C., Lin, S., & Li, W. (2015). Visual movement and mental-workload for pilot performance assessment. In *Engineering Psychology and Cognitive Ergonomics* (pp. 356-364). Springer International Publishing.

Vogl, J., Hayes, A., Aura, C., Milam, L., Temme, L., & St. Onge, P. (2020). *A case for a measurement of incremental workload.* (USAARL-TECH-TR--2020-038). U.S. Army Aeromedical Research Laboratory.

Kallus, K., Tropper, K., & Boucsein, W. (2011). The importance of motion cues in spatial disorientation training for VFR-pilots. *The International Journal of Aviation Psychology, 21*, 135-152.

Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., & Wickens, C. (2001). Comparison of expert and novice scan behaviors during VFR flight. *Engineering Psychology and Cognitive Ergonomics*. HCI 2016 International.

Kim, S., Prinzel, L., Kabar, D. B., Alexander, A. L., Stelzer, E. M., Kaufmann, K., & Veil, T. (2011). Multidimensional measure of display clutter and pilot performance for advanced head-up display. *Aviation, Space, and Environmental Medicine, 82,* 1013-1022.

Li, W., Horn, A., Sun, Z., Zhang, J., & Braithwaite, G. (2020). Augmented visualization cues on primary flight display facilitating pilot's monitoring performance. *International Journal of Human-Computer Studies*, *135*, 102377.

Mansikka, H., Simola, P., Virtanen, K., Harris, D., & Oksama, L. (2016). Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. *Ergonomics, 59*(10), 1344-1352.

Mansikka, H., Virtanen, K., & Harris, D. (2019). Comparison of NASA-TLX scale, Modified Cooper-Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks. *Ergonomics*, *62*(2), 246-254.

Mansikka, H., Virtanen, K., & Harris, D. (2019). Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Transactions on Human-Machine Systems, 49*, 1-9.

Mansikka, H., Virtanen, K., Harris, D., & Simola, P. (2016). Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test. *Applied Ergonomics, 56*, 213-219.

Müller, S., Schreiter, K., Luckner, R., & Manzey, D. (2017). Manual flying and energy awareness. *Aviation Psychology and Applied Human Factors, 7,* 18-27.

Nittala, S., Elkin, C., Kiker, J., Meyer, R., Curro, J., Reiter, A., Xu, K., & Devabhaktuni, V. (2018). Pilot skill level and workload prediction for sliding-scale autonomy. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1166-1173.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In *Handbook of Perception and Human Performance. Volume II. Cognitive Processes and Performance.* (pp.41-42). Wiley and Sons.

Scannella, S., Peysakhovich, V., Ehrig, F., Lepron, E., & Dehais, F. (2018). Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft. *Human Factors*, *60*(7), 922-935.

Stanton, N. A., Plant, K. L., Roberts, A. P., Allison, C. K., & Harvey, C. (2018). The virtual landing pad: Facilitating rotary-wing landing operations in degraded visual environments. *Cognition, Technology & Work, 20,* 219-232.

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Sladien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology & Work*, 20(3), 351-365.

Wang, Z., Zheng, L., Lu, Y, & Fu, S. (2016). Physiological indices of pilots' abilities under varying task demands. *Aerospace Medicine and Human Performance, 87*(4), 375-381.

Westbrook, C., Anderson, R.O., & Pietrzak, P.E. (1966). *Handling qualities and pilot work load*. Air Force Flight Dynamics Lab - Control Criteria Branch.

Wickens, C. D. (1984). Processing Resources in Attention. In *Varieties of Attention*. P. R. and D. D.R. Orlando, FL, Academic Press, Orlando, FL.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors, 50*(3). 449-455.

Yang, J. H., Kennedy, Q., Sullivan, J., & Fricker, R. (2013). Pilot performance: Assessing how scan patterns & navigational assessments vary by flight expertise. *Aviation, Space, and Environmental Medicine, 84,* 116-124.

Yu, C., Wang, E., Li, W., & Braithwaite, G. (2014). Pilots' visual scan patterns and situation awareness in flight operations. *Aviation, Space, and Environmental Medicine, 85*(7), 708-714.

Zhang, X., Xue, H., Qu, X., & Li, T. (2017). *Can fixation frequency be used to assess pilots' mental workload during taxiing? In Lecture Notes in Computer Science* (pp. 76-84). Springer International Publishing.

Zheng, Y., Lu, Y., Yuwen, J., & Shan, F. (2019). Predicting workload experienced in a flight test by measuring workload in a flight simulator. *Aerospace Medicine and Human Performance, 90*(7), 618-623.

Zheng, Y., Lu, Y., Wang, Z., Huang, D., & Fu, S. (2015). Developing a measurement for task complexity in flight. *Aerospace Medicine and Human Performance, 86,* 698-704.

U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama

All of USAARL's science and technical information documents are available for download from the Defense Technical Information Center.

https://discover.dtic.mil/results/?q=USAARL

Army Futures Command
U.S. Army Medical Research and Development Command