Psychophysiological Assessment in Pilots Performing Challenging Simulated and Real Flight Maneuvers

Bernd Johannes; Stefanie Rothe; André Gens; Soeren Westphal; Katja Birkenfeld; Edwin Mulder; Jörn Rittweger; Carla Ledderhos

BACKGROUND: The objective assessment of psychophysiological arousal during challenging flight maneuvers is of great interest to aerospace medicine, but remains a challenging task. In the study presented here, a vector-methodological approach was used which integrates different psychophysiological variables, yielding an integral arousal index called the Psychophysiological Arousal Value (PAV).

METHODS: The arousal levels of 15 male pilots were assessed during predetermined, well-defined flight maneuvers performed under simulated and real flight conditions.

RESULTS: The physiological data, as expected, revealed inter- and intra-individual differences for the various measurement conditions. As indicated by the PAV, air-to-air refueling (AAR) turned out to be the most challenging task. In general, arousal levels were comparable between simulator and real flight conditions. However, a distinct difference was observed when the pilots were divided by instructors into two groups based on their proficiency in AAR with AWACS (AAR-Novices vs. AAR-Professionals). AAR-Novices had on average more than 2000 flight hours on other aircrafts. They showed higher arousal reactions to AAR in real flight (contact: PAV score 8.4 ± 0.37) than under simulator conditions (7.1 ± 0.30), whereas AAR-Professionals did not (8.5 ± 0.46 vs.8.8 ± 0.80).

DISCUSSION: The psychophysiological arousal value assessment was tested in field measurements, yielding quantifiable arousal differences between proficiency groups of pilots during simulated and real flight conditions. The method used in this study allows an evaluation of the psychophysiological cost during a certain flying performance and thus is possibly a valuable tool for objectively evaluating the actual skill status of pilots.

KEYWORDS: psychophysiological arousal assessment, air-to-air-refueling, AWACS.

Johannes B, Rothe S, Gens A, Westphal S, Kirkenfeld K, Mulder E, Rittweger J, Ledderhos C. *Psychophysiological assessment in pilots performing challenging simulated and real flight maneuvers*. Aerosp Med Hum Perform. 2017; 88(9):1–7.

Q1

Traditionally, workload[9,15,23] is assessed by three main types of measurement approaches: assessment of the objective parameters of the task, measurement of behavioral and physiological responses, and assessment of the subjective appraisal given by the performer. It is known that objective and subjective assessments are only weakly correlated with each other, and correlations between load design and physiological measurements are often lacking.[28] While all methods have their advantages and weaknesses, we will focus on objective methods for assessing psychophysiological arousal. To date, the objective in-flight assessment of arousal in aircraft pilots[26] remains a challenge in aviation medicine. While technical obstacles regarding the acquisition and processing of large physiological data sets have been overcome, the interpretation of physiological indicators recorded to assess mental stress remains a problem. There certainly is need for objective and scaled measures of arousal. Especially under extreme conditions and situations it is helpful to have not only subjective reports, but objective measures to assess the state that a military operator is in. This is both for the personnel's health and for the sake of safe mission planning.

From the Institute of Aerospace Medicine, German Aerospace Center (DLR), Cologne, Germany; the Medical Squadron, NAEW&CF E-3A Component, Geilenkirchen, Germany; and the German Air Force Center for Aerospace Medicine (GAF CAM), Fuerstenfeldbruck, Germany.

This manuscript was received for review in October 2016. It was accepted for publication in June 2017.

Address correspondence to: Bernd Johannes, Ph.D., Institute of Aerospace Medicine, German Aerospace Center (DLR), Linder Hoehe, Koln NRW D-51147, Germany; bernd.johannes@dlr.de.

DOI: https://doi.org/10.3357/AMHP.4782.2017

A substantial amount of psychophysiological studies[16,24,25] exists on the development and verification of physiological indices of arousal. Berntson[3–5] and Cacioppo[6–8] provided a model of an autonomic space for cardiac control, whereas Porges[22] highlighted the phylogenetic development of autonomic control. For the identification of independent sources, e.g., heart rate regulation, factor analytical approaches showed promising results. Backs[1,2] and Lenneman and Backs,[20] for instance, were successful in verifying independent factor structures to disentangle sympathetic and parasympathetic components of the "autonomic space" in ECGs and impedance cardiograms. It should be acknowledged, however, that humans intrinsically respond differently such that raw measurements such as heart rate or skin conductance cannot be directly compared among individuals as indicators of arousal. Hence, for our own research a statistical scaling approach was developed that allows an interindividually comparable arousal assessment. For field applications only, those measurements were included that can be robustly and reliably registered under field conditions: electrocardiogram, skin resistance, finger temperature, and the finger
Q2 pulse wave.[10,12,14] We have used the eigenvectors (a set of eigenvectors is the primary result of a factor analysis) of some large data sets that had previously been obtained (for details see Johannes and Gaillard[10]) to construct an "arousal space" from the different psychophysiological data measured. The orthogonal dimensions were considered as representations of the independent autonomic influences upon different target organs,[4] whereas the length (scalar) of the vector sum [referred to as the Psychophysiological Arousal Value (PAV)] served to quantify arousal. The determination of the so-called "Autonomic Response Pattern" (ARP)[10] allowed a pattern-specific normalization of the "arousal space," thus providing an interindividual comparability of the PAV. The assessment of ARP is based on the individual's responses to a psychological protocol that induces a series of mentally loading tasks and relaxing phases in between. The levels and the reactivity of different physiological parameters were summarized in profiles which could be repeatedly classified into five different ARP.[10,13]

Objective psychophysiological arousal assessment has the advantage that it is not dependent on the openness of the test subjects, measurements can be taken instantaneously and continuously and with a high degree of temporal resolution, and are not confounding the events taking place at the same time. In summary, the PAV thus allows online monitoring and intra- and interindividual comparison of responses to a series of short-term events such as different kinds of flight maneuvers.

After comprehensive validation of the method by the German Institute of Aerospace Medicine (DLR),[10] the study presented here was to assess the PAV under real flight conditions. The primary goal of the study[17–19] was to test and verify the PAV under defined flight conditions that evoke well-reported arousal effects. As a second goal, this study aimed to address whether the effects of a simulated flight are comparable to real flight conditions. The third goal was to test the predictability of real flight arousal based on standard baseline conditions.

METHODS

Subjects

In total, 15 male Caucasian AWACS pilots (average age 38 ± 6 yr, BMI 27 ± 3) volunteered for the study. All pilots were individually and extensively informed about the study by the flight surgeon and were provided with an exposition of the experiments scheduled before giving a written informed consent. The pilots had long-standing flight experience ($>$2000 h) on different airplanes and they had been assigned to fly the AWACS prior to study inclusion. Based on their air-to-air refueling (AAR) experience in AWACS aircraft, the commanders of the participating squadrons divided them into two classes of proficiency, i.e., AAR-Novices ($N = 5$), and AAR-Professionals ($N = 10$).

Equipment

The study focused on specific load during air-to-air refueling of an AWACS airplane. Herein the real flights were done with a modified heavy class E-3 Sentry aircraft, which is a modified Boeing 707 aircraft. The aircraft is equipped with an external airborne radar picket system called the Airborne Warning And Control System (AWACS). The AWACS was historically also mounted to other airplanes and on ground stations. All measurements were carried out with the HealthLab system, a polygraph produced by Koralewski Industrie Elektronik oHG, Hambühren, Germany. All sensors and measurement modules of the system were integrated either into a body vest, which was used during psychophysiological baseline diagnostics, or, in the case of simulator and real flight conditions, into a biker belt (see **Fig. 1**). The physiological raw data were transmitted by Bluetooth and stored on a Samsung tablet PC in real time. The PC featured a touch screen that was used by the investigators to mark each flight maneuver. This setup provided an excellent indoor telemetry and allowed the subject to move freely following the preparation. The baseline test software, the monitoring software, and the software for the analysis of the physiological data were provided by SpaceBit GmbH, Berlin, Germany.

Procedure

The study was conducted at the multinational Geilenkirchen-Teveren Air Base in Germany, which is the main operating base of the E-3A Component of NATO. The participating pilots had to undergo three different study phases: psychophysiological baseline diagnostics, a simulated flight, and a real flight. The simulator and real flight protocol included 22 different flight phases, which were finally merged into the following 6 classes indexed as "Normal Flight," "Normal Approach," "50 ft AAR," "Contact," "Precision Final," and "Landing" (in detail below).

The study was standardized to the greatest possible extent and was identical for the simulator and the real flights. The recorded data could be checked in real time by the researcher under all study conditions.

The baseline assessment was used to classify the individuals' ARP to psychological stressors. For this purpose, a screening method was used that had been previously developed and verified. The pilots underwent a psychophysiological calibration

Fig. 1. Measurement equipment during simulator and real flights.

procedure during which alternating states of mental load and relaxation are induced. During these states, electrocardiogram, peripheral skin resistance, finger skin temperature, rate and depth of respiration, and pulse transition time are recorded continuously. In each experimental phase, blood pressure was assessed, both continuously at the finger and oscillographically at the arm. Based on these data, a classification function was used to assign the pilots to one of five distinct groups of ARPs. Details of these methods are described elsewhere.[10,12] The baseline assessment provided the reference values for single channel measures and integrated PAV scores.

All AWACS pilots participating in this study completed the standard training program that included regular training flights in the E-3A component flight deck mission simulator from CAE Electronics, Montreal, Canada. Notably, cockpit design and handling of the simulator used in this study greatly resemble those of the real Boeing E-3A Sentry AWACS aircraft. Moreover, with its full-motion simulation and high-resolution panoramic view through the cockpit windows, the simulator provides realistic training possibilities for pilots.

The pilots, who were already familiar with the measuring device from the baseline assessment, were prepared for continuous monitoring immediately before flight training, which started around 09:00. Preparation took about 20 min. During the course of the flight (either simulator or real flight), the various flight maneuvers were indicated by the researcher and recorded, such that this information could later be assigned to the psychophysiological data. In order to ensure that the data were related to the appropriate flight maneuvers, the research team was informed by the instructor when another flight phase started. Though the order of maneuvers was dependent on the existing training level of the respective pilot and the kind of maneuvers flown, it was quite homogeneous between all simulator flights. The standard training flights were performed under normal weather conditions with an E-3 Sentry aircraft as part of the pilot's education and training program. Like the simulator flights, the real flights were strictly defined by the training program.

Each flight usually involved four to six pilots, of which one or two participated in the experiment. The instructors attempted to include all maneuvers of interest to the research team into the training of the pilots participating in the experiment. For two volunteers, two flights were required to achieve a full set of data. In flight, the researcher was seated in the "fifth seat" or, if it was occupied, in the front part of the cabin close to the cockpit. In this case, the cockpit door remained open. The researcher was familiar with instrument flight regulations (IFR) communication and was able to follow the flight phases listening to the communication within the cockpit and the communication between the cockpit and air traffic control by means of a head set connected to the aircraft system.

One aim of the study was to estimate the arousal level evoked by the AAR maneuver in comparison to other standard maneuvers. It can be assumed that already simply approaching another large aircraft constitutes an extraordinary psychological challenge for the pilots. The usual minimum air separation between aircraft is about 1000 ft. During the contact phase of AAR, this distance is reduced to about only 15 ft. The boom from the tanker aircraft, large and heavy as it is, passes the cockpit windows very closely. Unlike jet fighters, which are more or less pulled from the tanker during the AAR contact phase, the AWACS aircraft is heavy and has to be controlled manually during that phase. Prediction of control effects during manual control of AWACS aircraft involves time delays due to the enhanced moment of inertia, and is thus inherently difficult. This becomes especially demanding under turbulent weather conditions. In addition, information about the contact position, which can be affected by weather conditions, is only visually available for the pilot. Due to the limited airspace reserved for the AAR maneuver, the flight path regularly involved 180° turns.

The air-to-air refueling maneuver started with a first communication contact between the aircraft. The tanker crew took control when the AWACS aircraft entered the 3-mile range.

Upon approaching the tanker, the AWACS pilot first had to stabilize his position behind the tanker aircraft. After receiving clearance by the tanker, the AWACS aircraft further approached the tanker. The tanker's boom operator then actively inserted the fuel boom into the docking neck (connecting piece) on top of the cockpit of the AWACS. This entire phase from the moment of direct contact until disconnection is hereafter called "contact."

To ensure reliable psychophysiological measurements, the instructors and the research team agreed upon an AAR contact time of 3 to 5 min. The contact phase was terminated either after an automatic disconnection due to turbulence or after the regular measurement period upon request of the AWACS instructor. Upon disconnection, the AWACS aircraft returned to the 50-ft AAR position.

During the simulated and real flights, a 1-lead electrocardiogram (ECG), skin resistance, finger temperature [FT (°C)], and pulse wave were registered continuously. The ECG was sampled at 1000 Hz for the system's internal analysis and down sampled to 500 Hz for storage. The electrodes were of the standard single-use Ag/AgCl-ECG type (Kendall/Arbo H 124 SG Ø 24 mm, Typo HealthCare Deutschland GmbH, Neustadt, Q3 Germany). Pulse wave, skin resistance, and FT were measured using an integrated multiuse finger sensor placed on the tip of the little finger of the hand not used for controlling the aircraft/simulator. Pulse wave was measured by using photoplethysmography with infrared light. The data were sampled at 500 Hz. Skin conductance level [SCL (μS)] was calculated from the skin resistance measured between the finger sensor (dry Ag sensor) and the mass electrode of the ECG using a maximum of 10 μA constant DC, i.e., measuring voltage sampled with 25 Hz. FT was registered using an FS-03/M thermo-sensor at a sampling rate of 5 Hz.

For each flight phase, the mean and SD of the following measures were calculated for further statistical analyses. ECG was used to obtain heart period duration [HPD (ms)], and the root of mean successive square differences [RMSSD (ms)] between R-peaks as a robust measure of vagal heart control. Pulse wave was used to obtain the pulse transit time [PTT (ms)], calculated as the interval between R-peaks of the ECG and the highest slope of the first pulse wave front. During the baseline assessment, it was also possible to register blood pressure both continuously at the left middle finger (CNAP, CNSystems, Graz, Austria) and oscillographically (Mobil-O-Graph, Q4 I.E.M. GmbH, Stolberg, Germany) at the right arm.

Statistical Analysis

The data presented here were statistically analyzed using IBM SPSS Statistics version 20. The Linear Mixed Effect (LME) model applied for the comparison among flight phases included as fixed effects the flight type and the flight phase. The pilot ID was set as a random effect. Variances were allowed to differ among pilots and the LME models were optimized according to the Akaike information criterion.[21] A model was accepted if the residuals were normally distributed. The level for statistical significance was set to $\alpha = 0.05$. However, due to the low statistical power, tendencies in the results (i.e., with P-values < 0.1) will also be reported. A correlation analysis was run between baseline, simulator, and real flight values using the Pearson correlation coefficient r.

RESULTS

In this manuscript we focus on the integrated PAV score. However, the raw data are given in **Appendix A** and can be viewed online (https://doi.org/10.3357/amhp.4782sd.2017). The calculation of the PAV score is based on the individual autonomic response pattern. Four out of five ARPs were observed in our cohort of subjects. Most frequently (nine times), pilots were of ARP type 1 (autonomic stable "non-responders"), four times of ARP type 2 (skin conductance responder), one time of ARP type 3 (heart rate responder), and one time of ARP type 5 (blood pressure responder). There was no significant relationship between the autonomic response pattern and the classes of proficiency (cc = 0.590, $P = 0.238$). All mentally relaxing phases (during the baseline measurement) were averaged to retrieve a solid 'mentally unloaded' baseline value called "Baseline."

Q5 In the integral PAV (**Fig. 2**), the changes with respect to the baseline were highly significant both in the simulator (df: num 6, denum: 69,079, $F(6, 69) = 11.079$, $P < 0.001$) as well as during real flight (df: num 6, denum: 69,115, $F(6, 69) = 12.293$, $P < 0.001$). The PAV showed significant interactions between the protocol phases and the proficiency groups for the simulator

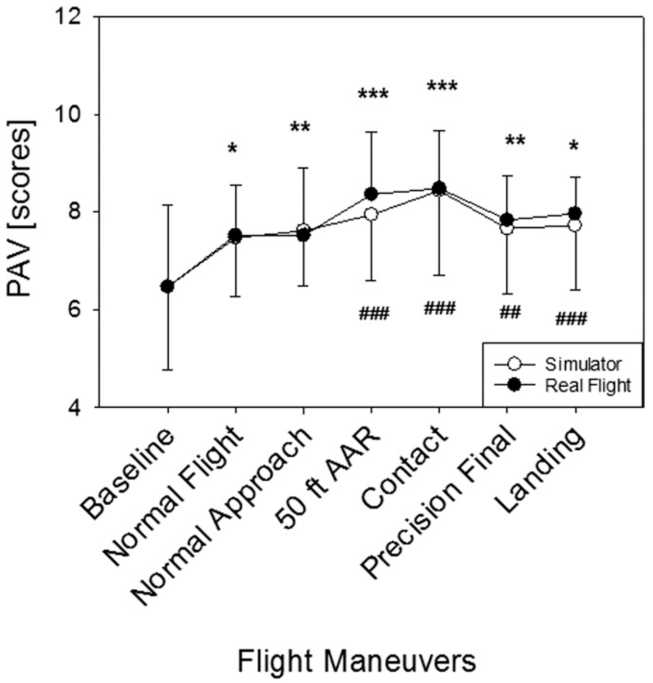

Fig. 2. Behavior of the integral PAV scores during the six phases of flight ("Normal Flight," "Normal Approach," "50 ft AAR," "Contact," "Precision Final," and "Landing") in comparison to the reference value called "Baseline" during two types of training: simulated flights (white circles) and real flights (black circles). The significance of differences from the baseline is given (real flights: $^{\#\#}P < 0.01$, $^{\#\#\#}P < 0.001$; simulator: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$). Q14

(df: num 6, denum: 69,079, $F(6, 69) = 2.937$, $P = 0.013$), but not for the real flights. A general comparison of the PAV in the two types of training provided no statistical differences.

The second aim of the study was to directly compare arousal under simulated and real flight conditions. Here, we focus on the effect of air-to-air refueling since this was the specific goal of the training program (**Fig. 3**). In addition, as expected, AAR also was the most challenging maneuver.

A significant difference with respect to the PAV was found in the AAR phases of the two types of training (df: num 1, denum: 32,419, $F = 5.376$, $P = 0.027$). Additionally, a tendency was found for the interaction between the training type and the proficiency classes (df: num 1, denum: 32,419, $F = 2.951$, $P = 0.095$). Separate analyses of both proficiency groups verified the difference between training types to be related to AAR-Novices. The AAR-Novices showed a significant difference in PAV scores between the simulator and real flights (df: num 1, denum: 13, $F = 4.894951$, $P = 0.045$), whereas the AAR-Professionals did not.

As part of the second aim of the present study, the workload of AAR was compared to the workload of landing maneuvers. **Fig. 4** depicts the differences between simulations and real flight conditions for the AAR-Novices and the AAR-Professionals.

In the group of AAR-Novices, the PAV tended to be higher during real flights as compared to simulator flights (df: num 1, denum: 30,758, $F = 3.778$, $P = 0.061$). In the group of AAR-Professionals, this was not the case. A tendency was also found for the threefold interaction between the loading effect of maneuvers, the training type, and the proficiency groups (df: num 1, denum: 31,516, $F = 2.900$, $P = 0.098$). Overall, distinct data in PAV between both training types showed a tendency toward significant differences in AAR-Novices (df: num 1, denum: 12, $F = 3.28$, $P = 0.095$), whereas no differences could be obtained in the group of AAR-Professionals.

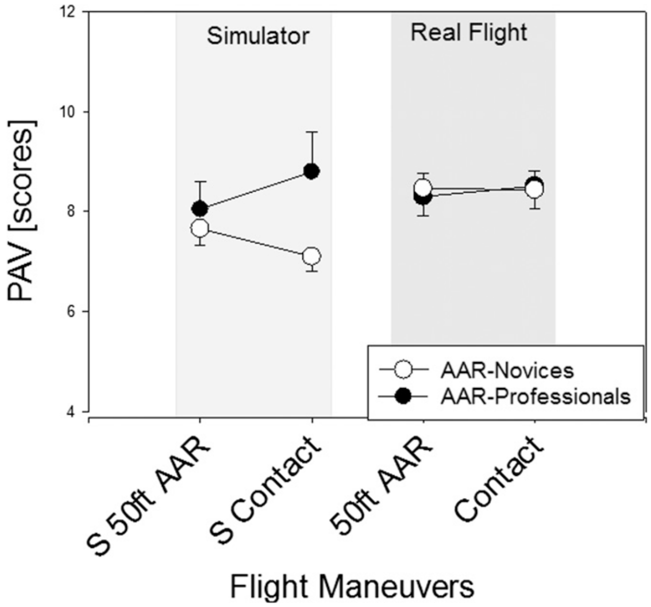

Fig. 3. Comparison of load levels of AAR phases during simulator and real flights. The behavior of the PAV indicated different reactions of the two proficiency groups in the two AAR phases (approach vs. contact phase) during simulator flights, but not in real flights. The white circles represent AAR-Novices and the black circles represent AAR-Professionals.

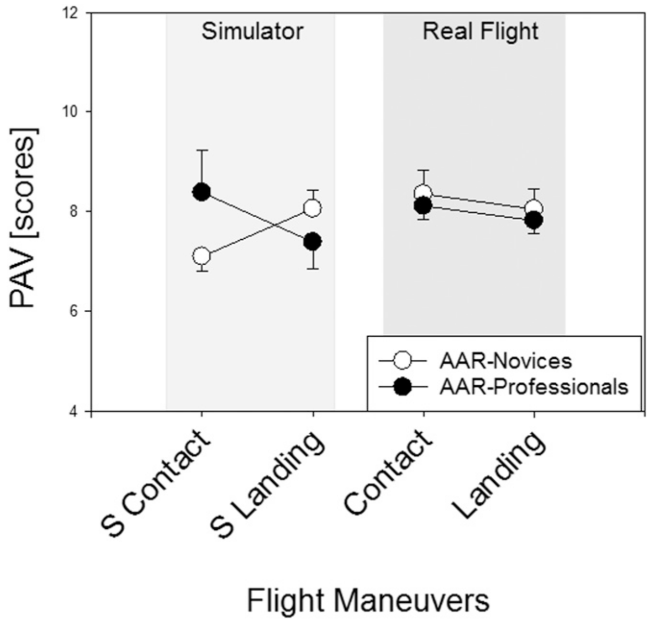

Fig. 4. Comparison of load levels (PAV) during the contact phase of AAR and the landing phase. The scores indicated different reactions of the proficiency groups during the contact phase of AAR and the landing procedures in the simulator and real flights. The white circles represent AAR-Novices and the black circles represent AAR-Professionals.

When both training types were analyzed together, no general difference was found in the PAV of the AAR contact phase and the landing maneuver. For the AAR-Novices, the landing maneuvers were equally loading as the AAR maneuver. For the AAR-Professionals, however, the AAR contact evoked significantly higher load levels than the landing maneuvers (df: num 1, denum: 22, 273, $F = 4801$, $P = 0.039$). In the simulator, the landing, as compared to the AAR maneuver, resulted in higher PAV scores in the AAR-Novices, whereas the opposite was found for the AAR-Professionals.

In general, the correlation analyses provided no significant predictive value of the baseline scores, neither for the simulator nor for the real flight measures, nor for single parameters, nor for the integrated PAV scores. Significant correlations between simulator and real flight data were found for single parameters (see **Appendix B**, which can be viewed online at https://doi.org/10.3357/amhp.4782sd.2017).

DISCUSSION

Acceptable psychophysiological costs are one of the basic conditions that determine the capacity of an individual to cope with and react to unexpected events and situations. As such, the assessment of psychophysiological arousal level values in pilots in combination with actual flight performance would be a potent tool since it can be used for evaluating pilot training status and progress during the training program. During active coping situations the chances to develop a higher level of arousal are high. This is exactly the trade-off we are using to investigate the level of "proficiency." We assume the more one person acts professionally in a certain operation, the less is his level of arousal. Methodologically we verified that the mobile psychophysiological measurement system HealthLab can be used successfully under standard flight conditions. The system was anecdotally nonobtrusive to the pilots and the scientific monitoring procedure using telemetric data transmission worked reliably. The physiological measurements taken in flight were of good quality and the selected statistical measurement parameters were robust enough for semiautomated analyses. More importantly, the method, which integrated various correlates of autonomous activation in different physiological measures into one integral value (PAV), provided plausible results. Flight phases that were commonly known to be more challenging to the pilots were indeed reflected by higher PAVs, indicating the validity of the methodology. Between proficiency groups, the PAV provided significant differences or interactions despite the limited number of pilots tested and the large interindividual variability in the underlying physiological raw data. The pilots in our study were classified into five ARP by a validated baseline screening method. Most of the pilots ($N = 9$) were classified as "non-responders," which is typical of specific, highly selected subject groups, such as pilots and rangers[27] or astronauts.[11] Six of the pilots showed a higher responsiveness to mentally loading tasks with various autonomic response patterns. These pattern groups differ significantly in level and reactivity magnitudes of the underlying physiological data.[10]

From an operational point of view, the comparison of the loading effect of training flights in the simulator with that of real flights was the main interest. Overall, the mean PAV levels were comparable between simulator and real flight conditions. Dividing the pilots according to proficiency revealed, however, that AAR-Novices showed noticeably higher PAVs, particularly with regard to air-to-air refueling in real flights as compared to simulator flights (Fig. 4). This finding is of special importance if one considers that even the AAR-Novices already had, on average, more than 2000 flight hours of general flying practice. They only were AAR-Novices with regard to the AWACS aircraft with its specific aerodynamic characteristics. Hence, the objective evaluation not only of flying performance, but, in particular, of the associated psychophysiological cost is a potential potent tool for objectively evaluating the training status and progress of AAR-Novices on their way to becoming AAR-Professionals.

A second operational aim was to objectively assess whether air-to-air refueling constitutes a significantly more demanding mental task than a landing maneuver, which is anecdotally reported and now quantitatively supported by our data. For the AAR-Novices, the landing maneuver was still similarly demanding.

A third aim was to analyze the predictability of real flight arousal based on standardized baseline measurements. There were notable correlations between simulator data and the data obtained in real flight (see Appendix B, Table BI, BII, and BIII, which can be viewed online at https://doi.org/10.3357/amhp.4782sd.2017) for single parameters, but not for the integrated PAV. This could be understood as an effect of the interindividual differences of the raw parameters providing correlations among situations, whereas the PAV neglects these individual features.

All in all, the objective assessment of psychophysiological workload developed by Johannes and Gaillard[10] was successfully applied under real flight conditions in the present study. Further research has to enhance the statistical power of single findings by increasing the sample size. The successful application of the nonobtrusive methodology and the semiautomated data analysis should make this feasible. Altogether, this method appears to be a promising approach for an objective and quantitative in-flight assessment of arousal.

ACKNOWLEDGMENTS

First of all, we would like to thank all participating pilots and flight engineers for their cooperation and their patience with the research team. In addition, we would like to thank the NAEW&C Force Command (Mons, Belgium) as well as Brigadier General S. Schmidt, former commander of the NATO airbase, Dr. E. Rödig, former Surgeon General of the German Air Force, and the commanders of the squadrons, especially Lt. Col. H. Grunwald, and Lt. Col. W. Ronneberger, head of the Simulation Center, for their support.

This work was conducted as a contract research project of the Bundeswehr Medical Service.

Authors and affiliations: Bernd Johannes, Ph.D., Dipl. Psych., Edwin Mulder, Ph.D., and Jörn Rittweger, M.D., Institute of Aerospace Medicine, German Aerospace Center (DLR), Cologne, Germany; Stefanie Rothe, M.D., Flight Surgeon, Medical Squadron, NAEW&CF E-3A Component, Geilenkirchen, Germany; and André Gens, Graduate Biomedical Engineer, Soeren Westphal, M.D., Katja Birkenfeld, M.D., and Carla Ledderhos, M.D., Ph.D., German Air Force Center for Aerospace Medicine, (GF CAM), Fuerstenfeldbruck, Germany;

REFERENCES

1. Backs RW. A comparison of factor analytic methods of obtaining cardiovascular autonomic components for the assessment of mental workload. Ergonomics. 1998; 41(5):733–745. Q6
2. Backs RW. An autonomic space approach to the psychophysiological assessment of mental workload. In: Hancock PA, Desmond PA, editors. Stress, workload and fatigue. London: Lawrence Erlbaum Associates; 2001:279–289.
3. Berntson GG, Cacioppo JT, Grossman P. Whither vagal tone. Biol Psychol. 2007; 74(2):295–300.
4. Berntson GG, Cacioppo JT, Quigley KS. Autonomic determinism: the modes of autonomic control, the doctrine of autonomic space, and the laws of autonomic constraint. Psychol Rev. 1991; 98(4):459–487.
5. Berntson GG, Cacioppo JT, Quigley KS. Autonomic cardiac control. I. Estimation and validation from pharmacological blockades. Psychophysiology. 1994; 31(6):572–585.
6. Cacioppo JT, Berntson GG, Binkley PF, Quigley KS, Uchino BN, Fieldstone A. Autonomic cardiac control. II. Noninvasive indices and basal response as revealed by autonomic blockades. Psychophysiology. 1994; 31(6):586–598.
7. Cacioppo JT, Tassinary LG, Berntson GG. Psychophysiological Science. Interdisciplinary approaches to classic questions about the mind. In: Cacioppo JT, Tassinary LG, Berntson GG, editors. Handbook of psychophysiology, 3rd ed. New York: Cambridge University Press; 2007:1–16.

8. Cacioppo JT, Uchino BN, Berntson GG. Individual differences in the autonomic origins of heart rate reactivity: the psychometrics of respiratory sinus arrhythmia and preejection period. Psychophysiology. 1994; 31(4):412–419.
9. Gaillard AWK, Wientjes CJE. Mental load and workstress as two types of energy mobilization. Work Stress. 1994; 8(2):141–152. **Q7**
10. Johannes B, Gaillard AWK. A methodology to compensate for individual differences in psychophysiological assessment. Biol Psychol. 2014; 96: 77–85.
11. Johannes B, Salnitski VP. Integration of different autonomic measures into common indicators of "psychophysiological costs". In: Goeters KM, editor. Aviation psychology: practice and research. Aldershot, UK: Ashgate; 2004:327–342.
12. Johannes B, Salnitski VP, Soll H, Rauch M, Hoermann HJ. De-individualized psychophysiological strain assessment during a flight simulation test - validation of a space methodology. Acta Astronaut. 2008; 63(7–10):791–799.
13. Johannes B, Salnitski VP, Thieme K, Kirsch K. Differences in the autonomic reactivity pattern to psychological load in patients with hypertension and rheumatic diseases. Aviakosm Ekolog Med. 2003; 37(1):28–42.
14. Johannes B, Wittels P, Enne R, Eisinger G, Castro C, et al. Non-linear function model of voice pitch dependency on physical and mental load. Eur J Appl Physiol. 2007; 101(3):267–276.
15. Kramer AF. Physiological metrics of mental workload: a review of recent progress. In: Damos DL, editor. Multiple-task performance. London: Taylor & Francis; 1991:279–328. **Q8**
16. Kramer AF, Sirevaag EJ, Braune R. A psychophysiological assessment of operator workload during simulated flight missions. Hum Factors. 1987; 29(2):145–160. **Q9**
17. Ledderhos C, Gens A, Johannes B. Zur Problematik der Messung der psychophysiologischen Beanspruchung von Luftfahrzeugführern während des realen Flugbetriebes. Wehrmed Mschr. 2008; 52(8): 251–255.
18. Ledderhos C, Rothe S, Gens A, Johannes B. Objective psychophysiological strain assessment of AWACS crews in simulators and real flight operations. Annual Military Scientific Research Report 2009. Bonn (Germany): Federal Ministry of Defence; 2010:76–77.
19. Ledderhos C, Rothe S, Johannes B. Psychophysiological strain assessment in real flight - a new approach. Brussels (Belgium): NATO; 2009. Report on NATO Headquarters, 13.07.2009.
20. Lenneman JK, Backs RW. The validity of factor analytical derived cardiac autonomic component for mental workload assessment. In: Backs RW, Bouscein W, editors. Engineering psychophysiology. Mahwah (NJ): Lawrence Erlbaum Associates; 2000:161–175.
21. Pinheiro JC, Bates DM. Mixed effects models in S and S-plus. New York: Springer; 2000.
22. Porges SW. The polyvagal theory: phylogenetic substrates of a social nervous system. Int J Psychophysiol. 2001; 42(2):123–146.
23. Sirevaag EJ, Kramer AF, Wickens CD, Reisweber M, Strayer DL, Grenell JF. Assessment of pilot performance and workload in rotary wing helicopters. Ergonomics. 1993; 36(9):1121–1140. **Q10**
24. Veltman JA, Gaillard AWK. Physiological indices of workload in a simulated flight task. Biol Psychol. 1996; 42(3):323–342. **Q11**
25. Wierwille WW. Physiological measures of aircrew mental workload. Hum Factors. 1979; 21(5):575–593. **Q12**
26. Wilson GF, Eggemeier FT. Psychophysiological assessment of workload in multi-task environments. In: Damos DL, editor. Multiple-task performance. London: Taylor & Francis; 1991:329–360. **Q13**
27. Wittels P, Johannes B, Enne R, Kirsch K, Gunga H. Voice monitoring to measure emotional load during short-term stress. Eur J Appl Physiol. 2002; 87(3):278–282.
28. Yeh Y, Wickens CD. Dissociation of performance and subjective measures of workload. Hum Factors. 1988; 30(1):111–120.

APPENDIX A. PHYSIOLOGICAL RAW DATA

Fig. A1 presents the single measures during baseline and flight phases. Significant changes from the baseline were observed for the HPD in both during the simulator flights (df: num 6, denum: 69,497, $F = 25.996$, $P < 0.001$) and the real flights (df: num 6, denum: 70,169, $F = 13.410$, $P < 0.001$). In all cases of the Linear Mixed Effect Model (LME), also below, the residuals were normally distributed. Heart period duration (HPD) is the interbeat interval, the interval between two R-spikes of the continuous ECG. In clinical applications heart rate is likely more appropriate. However, heart rate is the inverse function of the HPD that would provide confounding nonlinear influences on the factor structures used in this manuscript.

The pulse transit time (PTT) did not change in the simulator, but did during the real flight (df: num 6, denum: 69,314, $F = 4.658$, $P = 0.001$). The interaction between the flight phases and the proficiency groups was not significant during

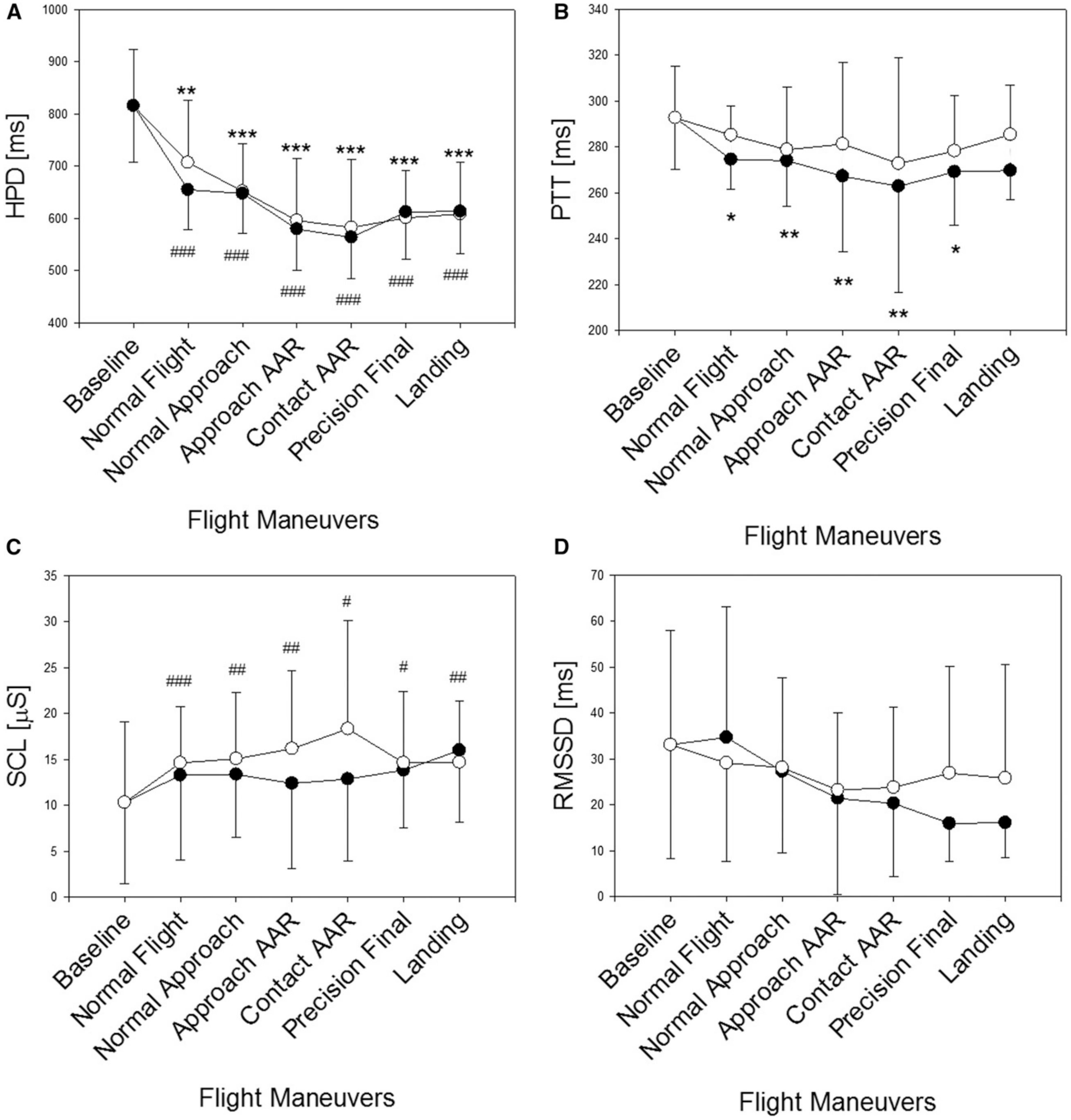

Fig. A1: Physiological results in six flight phases ("Normal Flight Activities," "Normal Approach Activities," "50-ft AAR," "Contact," "Precision Final," and "Touch and Go/ Landing") in comparison to a reference baseline "Baseline." The black circles represent data from the real flights; the white circles show data from simulated flights. The significant differences from the baseline are given (simulator: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; real flights: $^{\#}P < 0.05$, $^{\#\#}P < 0.01$, $^{\#\#\#}P < 0.001$).

Table BI. Correlations of HPD Scores Between Simulated and Real Flights.

		HPD_ NORMAPPROACH	HPD_50FT-TOAAR	HPD_-CONTACT	HPD_PREC-FINAL	HPD_ TOUCHANDGO
HPD_SNORM-APPROACH	Pearson Correlation	0.841	0.733	0.692	0.747	0.642
	Sig. (2-tailed)	0.000	0.003	0.006	0.005	0.018
HPD_S50FTTOAAR	Pearson Correlation	0.680	0.735	0.725	0.700	0.525
	Sig. (2-tailed)	0.005	0.003	0.003	0.011	0.065
HPD_SCONTACT	Pearson Correlation	0.610	0.703	0.690	0.642	0.511
	Sig. (2-tailed)	0.016	0.005	0.006	0.024	0.074
HPD_SPRECFINAL	Pearson Correlation	0.899	0.785	0.768	0.836	0.735
	Sig. (2-tailed)	0.000	0.001	0.001	0.001	0.004
HPD_STOUCHANDGO	Pearson Correlation	0.887	0.762	0.766	0.779	0.678
	Sig. (2-tailed)	0.000	0.002	0.001	0.003	0.011
	N	15	14	14	12	13

the simulator, but was during the real flights (df: num 6, denum: 69,314, $F = 3.228$, $P = 0.007$). Skin conductance level (SCL) changed during the simulated flights (df: num 6, denum: 69,082, $F = 4.302$, $P = 0.001$), but not during the real flights. The root of mean successive square differences (RMSSD) did not change significantly with respect to baseline values in either training condition.

Significant differences between the training types were not found in single data. However, the interaction of proficiency and training was significant for HPD (df: num 1, denum: 31,739, $F = 4.401$, $P = 0.044$), PTT (df: num 1, denum: 34,817, $F = 4.467$, $P = 0.042$), and SCL (df: num 1, denum: 31,385, $F = 10.496$, $P = 0.003$).

At the single parameter level, HPD showed no significant fixed effects. When analyzed separately, the professionals showed a significant interaction between the maneuvers and the training type (df: num 1, denum: 22,231, $F = 4582$, $P = 0.044$), supporting the impression that both groups reacted differently in simulator and real flights during the different maneuvers. The PTT analysis provided a near-significant interaction between proficiency classes and training types (df: num 1, denum: 33,974, $F = 3832$, $P = 0.059$) and for the beginners separately a near-significance of lower values during real flights (df: num 1, denum: 12, $F = 4614$, $P = 0.052$). In the SCL data no significant general fixed effect was found. However, the beginners, separately analyzed, showed significant differences between the training types (df: num 1, denum: 8056, $F = 10{,}017$, $P = 0.013$), the maneuvers (df: num 1, denum: 8056, $F = 5727$, $P = 0.043$), as well as a tendency toward significance for the interaction of maneuvers and training types (df: num 1, denum: 8182, $F = 4555$, $P = 0.065$). This interaction between maneuvers and training types was found to be significant for the professionals (df: num 1, denum: 22,191, $F = 5358$, $P = 0.030$).

APPENDIX B

A correlation analysis was performed to scrutinize the predictability of measures during the real flight based on measures during the simulator training. High correlations between simulator and real flight data were found for HPD (Table BI) and SCL (Table BIII). The integrated PAV scores (Table BII) showed tendencies for correlations, whereas no correlations were found for respiratory sinus arrhythmia (RMSSD) and finger temperature. The differing *N* indicates respectively the number of subjects having flown both maneuvers. The S following the variable name (e.g., HPD_S) indicates data from simulated flights. NormApproach stands for normal approach; 50FtToAAR and Contact the respective AAR phases; PrecFinal describes data from a precision final; and TouchAndGo indicates each kind of landing.

Table BII. Correlations of PAV Scores Between Simulated and Real Flights.

		PAV_NORM-APPROACH	PAV_ 50FTTOAAR	PAV_ CONTACT	PAV_ PRECFINAL	PAV_ TOUCHANDGO
PAV_ SNORMAPPROACH	Pearson Correlation	0.476	0.286	0.414	0.511	0.150
	Sig. (2-tailed)	0.073	0.321	0.142	0.090	0.625
PAV_ S50FTTOAAR	Pearson Correlation	0.453	0.327	0.445	0.558	0.335
	Sig. (2-tailed)	0.090	0.253	0.111	0.059	0.263
PAV_ SCONTACT	Pearson Correlation	0.511	0.372	0.521	0.553	0.371
	Sig. (2-tailed)	0.052	0.190	0.056	0.062	0.212
PAV_ SPRECFINAL	Pearson Correlation	0.521	0.359	0.530	0.507	0.237
	Sig. (2-tailed)	0.046	0.208	0.051	0.093	0.436
PAV_ STOUCHANDGO	Pearson Correlation	0.447	0.307	0.447	0.470	0.107
	Sig. (2-tailed)	0.095	0.285	0.109	0.123	0.728
	N	15	14	14	12	13

Table BIII. Correlations of SCL Scores Between Simulated and Real Flight.

		SCL_ NORMAPPROACH	SCL_ 50FTTOAAR	SCL_ CONTACT	SCL_ PRECFINAL	SCL_ TOUCHANDGO
SCL_ SNORMAPPROACH	Pearson Correlation	0.672	0.594	0.585	0.615	0.562
	Sig. (2-tailed)	0.006	0.025	0.028	0.033	0.045
SCL_ S50FTTOAAR	Pearson Correlation	0.774	0.776	0.771	0.793	0.808
	Sig. (2-tailed)	0.001	0.001	0.001	0.002	0.001
SCL_ SCONTACT	Pearson Correlation	0.721	0.708	0.723	0.741	0.803
	Sig. (2-tailed)	0.002	0.005	0.003	0.006	0.001
SCL_ SPRECFINAL	Pearson Correlation	0.749	0.675	0.661	0.710	0.692
	Sig. (2-tailed)	0.001	0.008	0.010	0.010	0.009
SCL_ STOUCHANDGO	Pearson Correlation	0.690	0.659	0.635	0.657	0.617
	Sig. (2-tailed)	0.004	0.010	0.015	0.020	0.025
	N	15	14	14	12	13

Author Query sheet–AMHP4782

Q1 : Author: I have added the missing references per our email conversation. Please check that they have been placed correctly.
Q2 : Author: Refs. 3-5, 6-8, 10-14, and 18-19 were re-organized and renumbered to be in alphabetical order per our style. Please check that citations were correctly transposed in the text.
Q3 : Author: Is Neustadt the correct location for Typo HealthCare? If not, please provide the correct city and country.
Q4 : Author: Is I.E.M. the correct manufacturer of the Mobil-O-Graph and is Stolberg, Germany, the correct location for them? If not, please provide the correct manufacturer and location (city and state or country).
Q5 : Author: Is this an appropriate place to cite Fig. 2? If not, please indicate where it should be cited.
Q6 : Reference 1 "Backs, 1998" is not cited in the text. Please add an in-text citation or delete the reference.
Q7 : Reference 9 "Gaillard, Wientjes, 1994" is not cited in the text. Please add an in-text citation or delete the reference.
Q8 : Reference 15 "Kramer, 1991" is not cited in the text. Please add an in-text citation or delete the reference.
Q9 : Reference 16 "Kramer, Sirevaag, Braune, 1987" is not cited in the text. Please add an in-text citation or delete the reference.
Q10 : Reference 23 "Sirevaag, Kramer, Wickens, Reisweber, Strayer, Grenell, 1993" is not cited in the text. Please add an in-text citation or delete the reference.
Q11 : Reference 24 "Veltman, Gaillard, 1996" is not cited in the text. Please add an in-text citation or delete the reference.
Q12 : Reference 25 "Wierwille, 1979" is not cited in the text. Please add an in-text citation or delete the reference.
Q13 : Reference 26 "Wilson, Eggemeier, 1991" is not cited in the text. Please add an in-text citation or delete the reference.
Q14 : Author: In Fig. 2, the single # P-value description was removed as there is no single # in the figure. Is this correct?