Operator's Reliability During Spacecraft Docking Training on Board Mir and ISS

Bernd Johannes; Sergey V. Bronnikov; Juri A. Bubeev; Tatyana I. Kotrovskaya; Daria V. Shastlivtseva; Sarah Piechowski; Hans-Juergen Hoermann; Jens Jordan

BACKGROUND: The experimental research "PILOT" on board the space stations aimed to assess cosmonauts' expectable reliability in a mission relevant operation, the manual docking of Soyuz or Progress spacecrafts on the space stations Mir and International Space Station (ISS), respectively.

METHOD: Therefore, a simulation of the docking of two space apparatuses was used for training and research. The methodological approach is described, taking into consideration the level of performance and the individual effort spent, the "psychophysiological costs". In three decades altogether 32 cosmonauts took part.

RESULTS: A significant increase of reliability was found from Mir (0.45 scores) to ISS missions (0.51). On ISS the reliability remained stable (0.50 ± 0.1).

DISCUSSION: Sal'nitskiĭ's model for the evaluation of operator's reliability was further developed and tested, which turned out to be sensitive as well as robust enough for a practical application in this critical operational task.

KEYWORDS: human reliability modeling, performance assessment, psychophysiological measurement, manual docking.

Johannes B, Bronnikov SV, Bubeev JA, Kotrovskaya TI, Shastlivtseva DV, Piechowski S, Hoermann H-J, Jordan J. *Operator's reliability during spacecraft docking training on board Mir and ISS*. Aerosp Med Hum Perform. 2021; 92(7):541–549.

"… the Soyuz craft carrying Britain's first European Space Agency astronaut, Tim Peake, as it prepared to dock with the International Space Station. A tense, last minute glitch with the Soyuz forced the crew to make an unusual manual approach to the orbiting outpost, but all turned out well."[24] Timothy Peake is the last European benefitting from the skills of his Russian crew mate Yuri Malenchenko to dock manually at the International Space Station (ISS). Answering the question "What was your scariest moment in space?", he responded that this docking maneuver was "the moment of greatest apprehension to me."[18] In 1997, the German cosmonaut Reinhold Ewald arrived at the Mir station after a manual docking maneuver performed by Soyuz Commander V. Zibliev [Ewald R. 2020, private communication]. Shortly after he had returned to Earth, a cargo craft crashed into the Mir module Spectr during a manual docking attempt.[7] As the first non-Russian, the European Space Agency astronaut Thomas Reiter successfully trained in the docking maneuver and served as Soyuz commander during return to Earth.[8] While several Europeans achieved this qualification, they never executed the maneuver. Manual docking can determine success or failure of an entire mission. Therefore, reliability of this operation should be maximized. Moreover, predictors for manual docking success should be identified.

Human Reliability Analysis (HRA) holds promise in this regard. With this approach, the nuclear industry develops methods to protect nuclear power plants from human errors. Bell and Holroyd[3] reviewed commonly applied methodologies which primarily calculate probabilities of critical errors in large plants or factories. Another approach focuses on operator performance under emergency conditions to gauge human reliability.[15,17,23] The authors applied a nuclear plant simulator and collected plant-specific and domain-specific human

From the German Aerospace Center (DLR), Institute of Aerospace Medicine, Cologne, Germany; the S.P Korolev Rocket and Space Corporation "Energia", Moscow, Russia; the Institute for Biomedical Problems (IBMP) of the Russian Academy of Sciences, Russian Federation State Research Center, Moscow, Russia; and the Department of Aerospace Medicine, University of Cologne, Germany.

This manuscript was received for review in August 2020. It was accepted for publication in March 2021.

Address correspondence to: Bernd Johannes, Ph.D., Institute of Aerospace Medicine, Linder Höhe, D-51147, Cologne, Germany; bernd_johannes@hotmail.com.

DOI: https://doi.org/10.3357/AMHP.5745.2021

performance data. Reliability was defined as the probability of successful performance of activities required for reliable system function. Similarly, we previously tested for performance predictors during manual docking.[14] Lager provided an overview on requirements and reliability research in aviation.[16]

In this paper, we apply an approach taking into consideration performance levels in preceding tasks and the required psychophysiological effort. Meanwhile, a large set of docking training data is available from spaceflights,[5,6,20] as well as from laboratory and space analog simulation studies, for example, isolation, bedrest, dry immersion, and Antarctica.[4] According to Sal'nitskiĭ, Dudukin, and Johannes,[21] we extend the former reliability definition: an operator's reliability is the probability he/she will fulfill a necessary operation with required quality and appropriate effort.

We previously assessed and evaluated "appropriate effort"[2,12,13] by combining psychophysiological parameters into a Psychophysiological Arousal Value (PAV). The methodology allows comparing different individuals, taking into account their actual autonomic response pattern (ARP). Psychophysiological parameters measuring multiple aspects of "workload", "emotional stress", and "activation"[9,19,25] have often been used as psychophysiological effort indicators. Heart rate (HR) and heart rate variability (HRV) have been analyzed in early studies. Eye blink rates, skin conductance,[27] and various electroencephalogram-derived parameters have also been proven useful. Voice frequency analyses as indicators of pilots'[26] or astronauts' emotional stress[11] is another promising approach for psychophysiological investigations. Voice parameters have a nonlinear relationship to physical and mental load, providing independent information about an operator's state compared to cardiovascular parameters.

Most of these studies estimated each measure's reliability and validity in relation to each other, environmental parameters, objective load assessments, subjective ratings (e.g., NASA Task Load Index), or performance data. These studies have proven the applicability, feasibility, and usefulness of psychophysiological measures under field conditions.

We previously integrated cardiovascular and performance measures into a common reliability index. The relationship between integrated objective indicators of performance, such as docking accuracy, as well as for the effort (herein PAV) were examined. Sal'nitskiĭ, Dudukin, and Johannes[21] integrated these measures using canonical correlation analysis. The method was successfully applied to data obtained in an isolation study (SFINCSS, Moscow, IBMP, 1999–2000). Thus, statistical methodology evaluating operator reliability changed from the analysis of single primary performance parameters and psychophysiological state to a more integrated approach. We suggest that this change permits a more comprehensive evaluation of operator's reliability and may improve prediction. The comparative analysis of integrated behavioral performance parameters and effort provided higher sensitivity of the psychophysiological effort parameters to inopportune and disturbing environmental conditions and their determining role in the evaluation of work reliability in human operators.[22] The effort indicator reacts immediately, whereas the performance still may maintain at a high level.

In another approach, performance data were separately analyzed for the different flight phases stabilization, final approach, and docking contact using factor models.[14] Psychophysiological data were combined into a PAV based on the cosmonaut's individual ARP.[12] Individual response pattern provides a more valid comparison of psychophysiological measures between persons as effort indicators. Combining the main factors of the psychophysiological datasets and the performance data separately provided two independently calculable scales for effort and performance, whereas the above-used canonical correlation requires both data sets for calculation.

After presenting previous approaches to assess operator reliability, we will focus on the actually used methodological approach in this Russian-German cooperation and the respective data of three experimental periods of onboard training of manual docking. The main aim of this paper was to determine and to present cosmonaut's reliability in this mission relevant skill.

METHODS

Subjects

From 1996 to 2001 on the Mir station, as well as 2008–2011 and 2015–2018 on the ISS, all 32 Russian cosmonauts (male, preflight mean weight 79.7 kg, 1.77 cm, 46 yr) underwent 3 preflight (−1 mo, −10 d, −3 d prior launch) and 3 postflight (+3 d, +10 d, +2 to 3 mo postlanding) experiments. The individual flight duration varied around 6 mo, ranging from 164 to 195 d in space. One subject remained 386 d on board. In flight, cosmonauts executed the experiment on Mir sporadically and on ISS on a regular monthly interval.

During the Mir epoch five Russian male cosmonauts participated actively in the experiment. Before the first data collection, the subjects had at least one training session with detailed instructions on experimental procedures. On ground the preparation phase was supported by the Russian investigators.

During the first ISS epoch (2008–2011) 12 and, during the second ISS period from 2015 to 2018, 16 crewmembers (15 Russian, 1 American) took part in the experiment. Two data sets had to be excluded for different reasons.

The method "PILOT" was created in 1987 for the investigation of cosmonauts' performance reliability in a simulated training task of hand-controlled approach and docking of a spacecraft (SC) "Soyuz" on the space stations "Mir" and "ISS" during long-term flights.

During the Mir period, the research simulator software was developed by Salnitski's group at the IBMP (mainly by Jury Shlykov). For the ISS periods, the software was provided by RSC Energia and was primarily used for regular docking training. Aiming at the development of EEG experiments, a new experimental docking simulator (6 df[10]) was created by DLR for the second ISS period. While visualizations differed between software, performance assessment was identical.[14]

Dynamic equivalence of the simulation to real docking maneuvers was verified for each simulator by cosmonauts to avoid negative training effects. During the first two periods, original

standard controls for the spacecraft were used. For the actual simulator and the numerous ground studies laboratory hand controls were developed by Koralewski. Functionally, these controls resemble the original controls. Psychophysiological parameters were registered using different generations of the Neurolab system (Neurolab-B, Neurolab-2000M, Neurolab-2010). Neurolab-B was produced by the Bulgarian Academy of Science in Sofia, Bulgaria, and both later generations were produced by Koralewski Industrie Elektronik oHG, Hambühren, Germany. For the first two generations, all sensors and measurement modules were integrated into body vests. The actual onboard polygraph is a small on-table application that can be integrated into a vest. The three polygraph generations featured some specific measurement channels, but were comparable in the main channels, described below.

Protocol

Subjects installed the equipment and applied the electrodes themselves following instructions provided on the computer screen. The central system block was fixed to a table in the central compartment of the Mir station and connected with the onboard power supply. After the system was booted, each step was guided by menus with illustrated and written instructions. During the experiment, instructions were mostly given acoustically through headsets from recordings with a native speaker. Cosmonauts ran 4–5 sessions on Mir, 3 tasks each; on ISS 6–12 sessions, 5 tasks each. The experiments on board Mir were run in the Russian language.

In a first experimental phase, the cosmonauts conducted the screening for their ARP. Since the start of these experimental series in space, this part was a separate experiment (Regulation). The docking experiment (PILOT) followed immediately. During the second ISS period both experimental phases were integrated into one experimental procedure. The ARP was needed to determine autonomic states and to apply type-specific integration functions for psychophysiological measures. Similar to a calibration, the measurement was obtained prior to docking training using the same equipment and the same sensors. The participants gave their informed consent and the study was approved by the IRB of the international ISS authorities.

Measurements

During simulated docking 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. 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. The pulse wave was measured by infrared photo plethysmography sampled with 500 Hz. Skin conductance level [SCL (μS)] was calculated from the skin resistance measured between the finger sensor (dry Ag sensor) and the ECG mass electrode using a maximum of 10 μA constant DC, i.e., measuring voltage sampled with 25 Hz. The latest generation finger sensor featured two integrated Ag strips. The FT was registered using an FS-03/M thermo-sensor at a sampling rate of 5 Hz. For each experimental phase, the individual mean and SD of the following measures were calculated for further statistical analyses. The ECG was used to obtain the heart period duration [HPD (ms)] and the root of mean successive square differences [RMSSD (ms)] between R-peaks as an estimate of vagal heart control. Pulse transit time [PTT (ms)] was calculated as the interval between R-peak of the ECG and the highest slope of the first pulse wave front.

Voice was registered with an 8-kHz sampling rate, sufficient for fundamental frequency assessment.[11] A commercially available head-set microphone (Sennheiser) was used. Because registered sound samples had to be considered as running speech, more sophisticated methods similar to utterance analyses could not be applied. Instead, we applied a robust approach based on voice pitch mode (f0m).

Reliability Modeling

Sal'nitskiĭ's[21] canonical correlation analysis approach determines the best variance explaining model (Eq. 1) of two integrated parameters, L_u (docking accuracy) and L_s (PAV), one for each of both independent data sets which are assumed to together represent one system—here, the man-machine system.

$$\begin{aligned} L_u(Y_n) &= b_0 + b_1Y_1 + b_2Y_2 + \ldots + b_nY_n \leftrightarrow L_s(X_n) \\ &= a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n \end{aligned} \qquad \text{Eq. 1}$$

Where Y_n is performance data, X_n is psychophysiological data, and a_n and b_n are weight values.

$$PR = \lambda_1 P(1-\lambda_2 C) \qquad \text{Eq. 2}$$

Individual Professional Reliability estimation (PR; Eq. 2), was based on normalized values of $L_u \rightarrow \lambda_1$ and $L_s \rightarrow \lambda_2$ between 0 and 1, providing an integrated value PR between 0 and 1. Sal'nitskiĭ's understanding of Eq. 2 PR was: high performance with acceptable effort predicts high reliability of the operation. Richter and colleagues[19] used the ratio (performance/effort) for the same purpose, but used the single psychophysiological channels phasic skin conductance response and the 0.1-Hz component of the heart rate variability.

Sal'nitskiĭ, in general, compared the current physiological indicator reduced by the minimum value divided by the range (max–min) of the available data. Reliability was then calculated as the product of the normalized performance value with the difference of the normalized physiological value from one, providing reliability scores ranging from 0 to 1.

$$\text{Effort} = \frac{(\text{arousal}_{\text{actual}} - \text{arousal}_{\text{min}})}{(\text{arousal}_{\text{max}} - \text{arousal}_{\text{min}})} \qquad \text{Eq. 3}$$

$$\text{Reliability}_{\text{Salnitski}} = \text{Performance}\,(1 - \text{Effort}) \qquad \text{Eq. 4}$$

Richter instead used the direct quote of performance and psychophysiological score.

$$\text{Reliability}_{\text{Richter}} = \text{Performance: physiological Activity Score} \quad \text{Eq. 5}$$

The results of these "classical" methods will be presented with some new developments, suggested by Sal'nitskiĭ. In a first new step, the performance value for a single training flight was created, averaging it with the previous two flights. Using a moving window, an overall session score was also calculated for a whole training session. A second step calculated an averaged value from the last three sessions respectively, providing a moving window (MW) averaged score.

$$\text{MW\_Reliability}_{\text{Flight}k} = \frac{(\sum_{n=k-3}^{k} \text{Flight Reliability}_n)}{3} \quad \text{Eq. 6A}$$

This is a simple splining approach to assess a value for each flight using a time series moving window for averaging.

$$\text{MW\_Reliability}_{\text{Session}k} = \frac{(\sum_{n=k-3}^{k} \text{Session Reliability}_n)}{3} \quad \text{Eq. 6B}$$

For the single session reliability, the single flights were sorted before by reliability, thus excluding the worst two.

Based on these session scores an Expected Reliability Value was calculated. A simple approach was the linear extrapolation by regression analysis:

$$\text{Reliability}_{\text{Regression}_{n+1}} = \beta + \beta_{\text{Repetition}_{n+1}} + \varepsilon_n \quad \text{Eq. 7}$$

An alternative measure is conveyed by a Bayesian hypothesis test;[1] we provide an example in the online figures (see Fig. B, https//doi.org/10.3357.AMHP.5745sd.2021).

Statistical Analysis

We applied IBM SPSS Statistics version 22. The Linear Mixed Effect (LME) model for the comparison between mission phases and docking tasks included the periods in space as an additional fixed effect. The cosmonauts' ID was set as a random effect. Variances were allowed to differ among cosmonauts and the LME models were optimized according to the Akaike information criterion. The numerator as well as the denominator degree of freedom for the respective *F*-values are presented. A model was accepted if the residuals were normally distributed. The level for statistical significance was set to $\alpha = 0.05$.

RESULTS

We present first the data on a selected single psychophysiological parameter level to provide evidence of their validity (**Fig. 1**). The complete results of the single psychophysiological data are presented in the supplemental material (online, https//doi.org/10.3357.AMHP.5745sd.2021). We included in this manuscript only the results of the PAV and HPD.

The HPD differed significantly between the mission phases [$F(2, 5382.303) = 589.293$, $P < 0.001$], showing longer HPDs in space and shorter HPDs postflight. The experimental phases were significantly different [$F(16, 5378.143) = 4.221$, $P < 0.001$], covarying with the loading character of the protocol. An interaction between the periods in space and the mission phases [$F(4, 5382.091) = 50.091$, $P < 0.001$] indicates that the protocol impacted the HPD differently during the three mission phases. There was no significant difference between the periods in space. These statistical results confirm the concordance of the

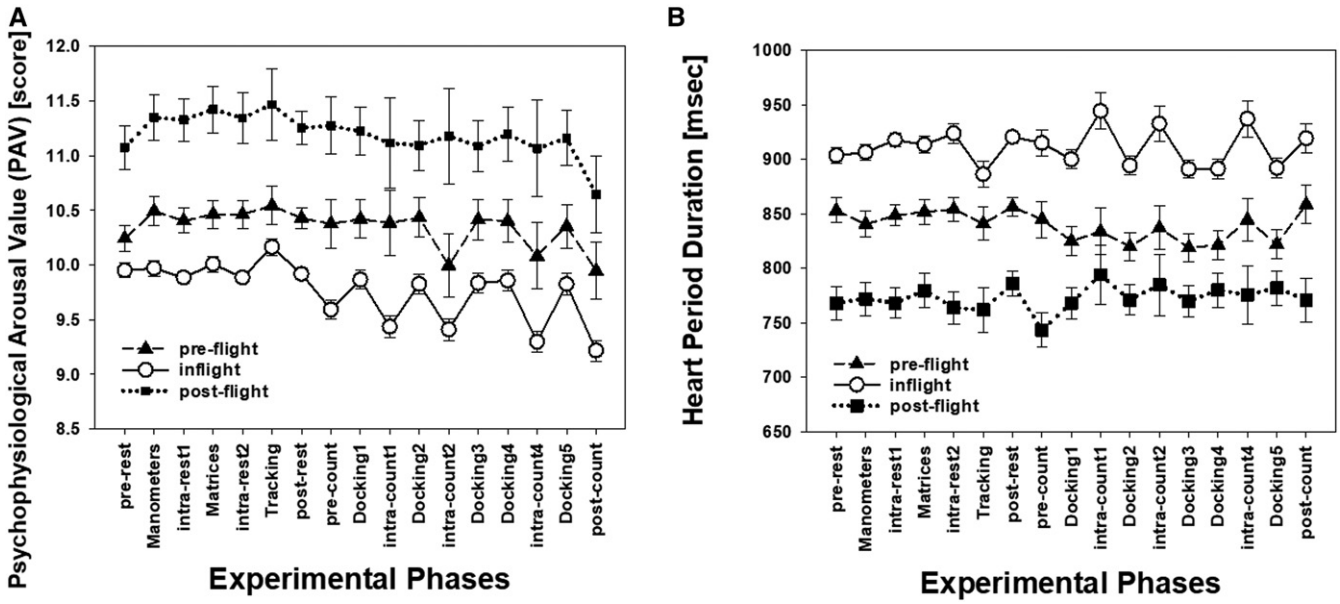

Fig. 1. A) Psychophysiological arousal value (PAV) and B) averaged HPD values (ms) over experimental phases (ARP screening + docking training) compared between mission phases.

HPD data with well-known literature findings. Assessing the effort with only one psychophysiological parameter leads to different results.

The second step was to investigate the relationship of these effort indicators with the performance results. This relationship is at least three-dimensional. **Fig. 2** presents the inverted U-surfaces for the relationships of PAV (fig. 2A–C) and HPD (fig. 2D–F) with Performance.

There are clear changes between mission phases, indicating a change from one-modal to a multimodal distribution (Figs. 2A–C). However, averaging these data by putting combinations of effort-performance pairs into percentage classes (transformation of 3D data into 2D data) provide (Fig. 2G) significant differences of the experiment phases between periods in space [$F(19, 1140.000) = 8.507$, $P < 0.001$], but no general differences between mission phases (Fig. 2H). This information is lost in the 2D representation.

In a third step the effort and reliability scores were analyzed. **Figs. 3A and B** demonstrate that the numerical value of the reliability score depends on the chosen psychophysiological parameter and is relatively constant within one training session. However, large differences can be observed when comparing the indicators across the mission phases (Fig. 3C). The reliability scores based on heart rate or skin conductance showed opposite reactions in flight, which should be further discussed.

The different reliability scores were significantly different in their mean [$F(4, 6979) = 70.836$, $P < 0.001$] and showed a significant interaction with the mission phases [$F(8, 6979) = 74.639$,

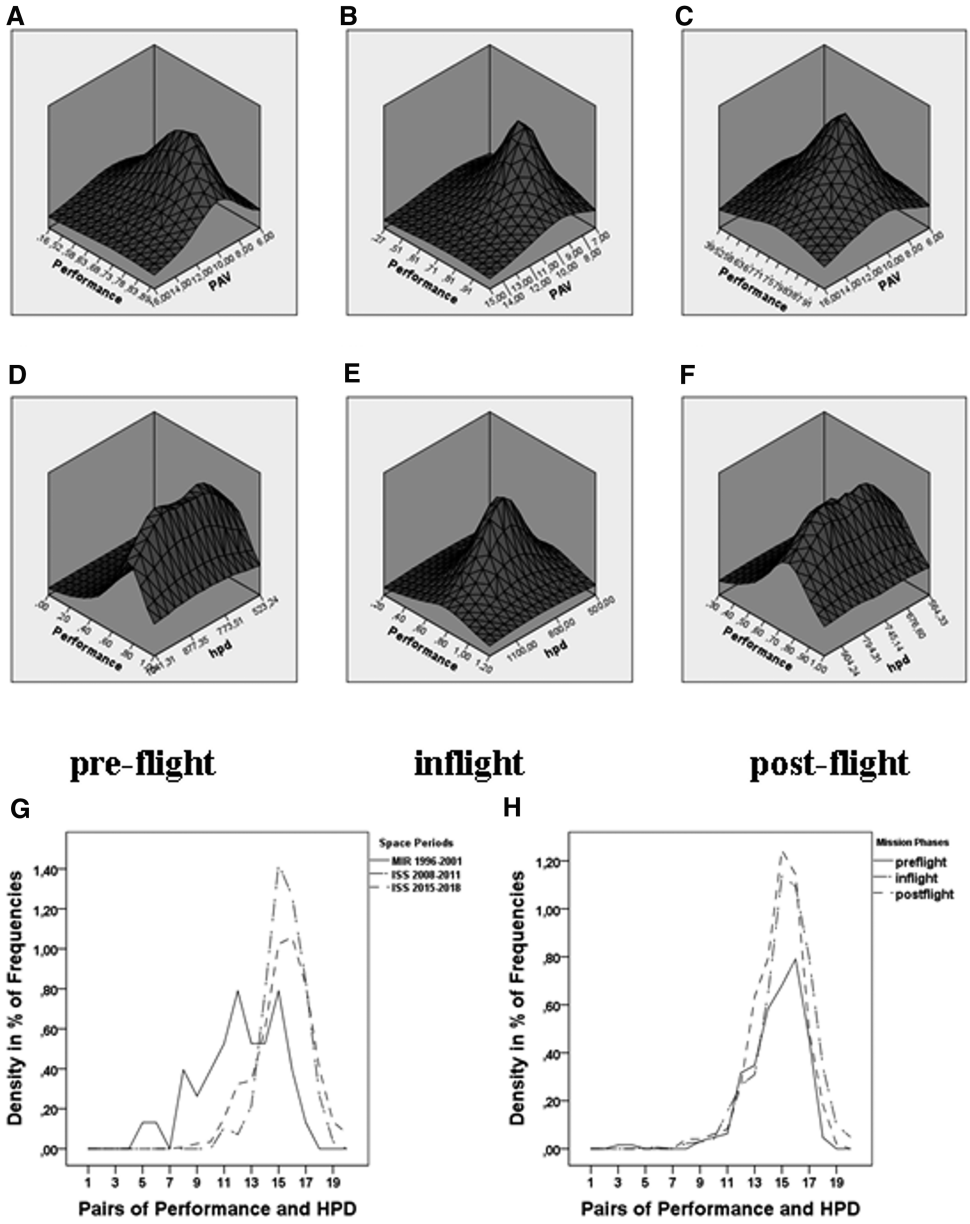

Fig. 2. Density (z-axis) surface of the relations between PAV scores (A-C) and HPD (ms)(D-F) as effort indicators and docking performance (score) during mission phases. Frequencies of effort-performance combinations G) between space periods and H) between mission phases.

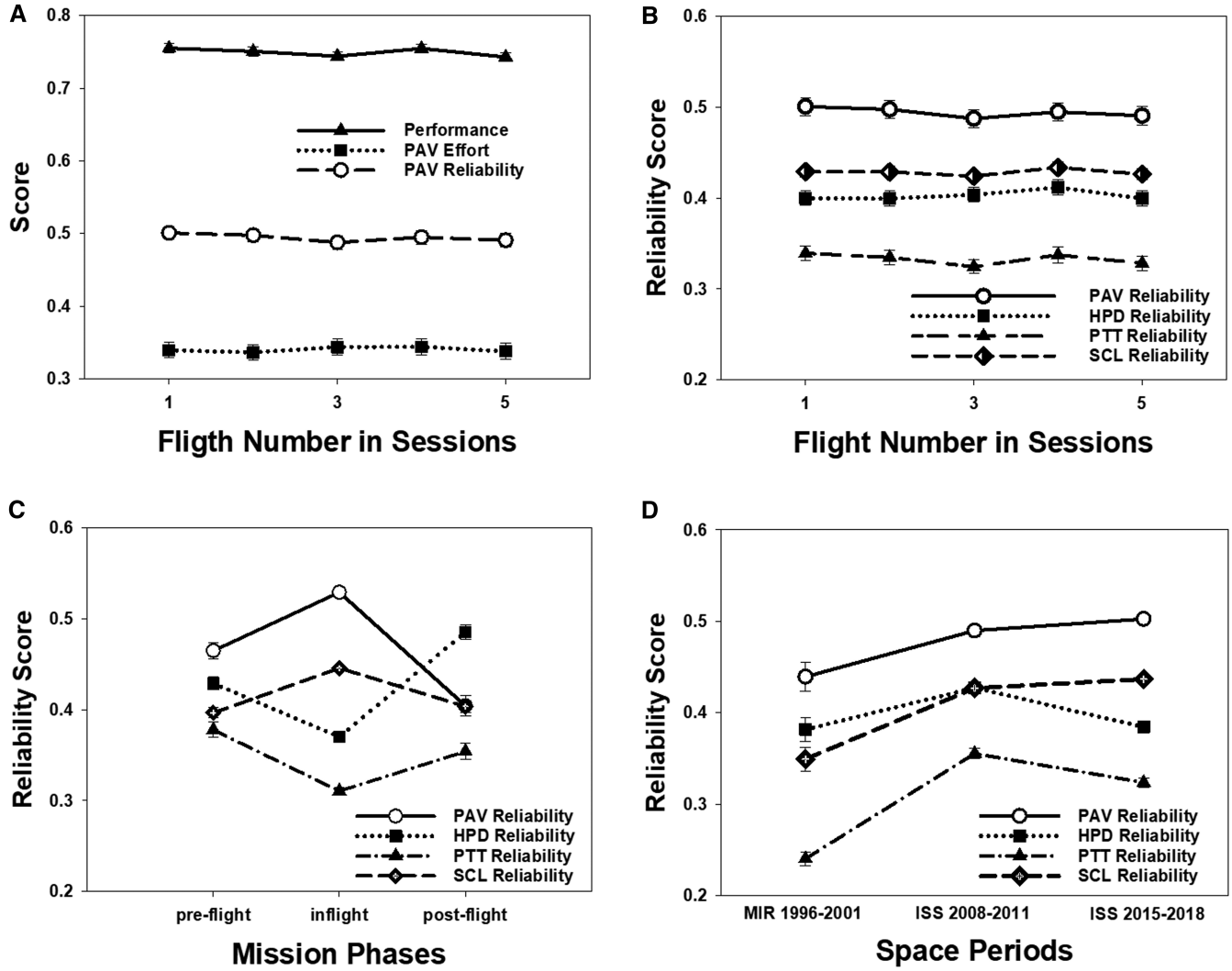

Fig. 3. Sal'nitskiĭ's reliability scores A) based on the integration of effort and performance, B) based on the single physiological parameters HPD, PTT, and SCL, C) over mission phases, and D) in flight between space periods.

$P < 0.001$], as well as with the periods in space [$F(8, 6979) = 3.947$, $P < 0.001$], underlining the different response to both factors.

To improve the predictive value of training results, splining could help to eliminate nonrelevant situational influences. The simplest way is a moving average of the single training flight results. In the presented data, an averaging over three training flights was applied. Additionally, reliability scores of whole training sessions were splined, providing a clearer development pattern over time (**Fig. 4**); three sessions were averaged.

Based on these training results, one can predict the expected result following training or during real docking. Linear regression predictions were applied for performance and effort as well as for integrating reliability scores. The regression was run based on the time series excluding the last two training flights or sessions. These two scores served for the verification of the prediction. There were no differences between predicted and observed scores (**Fig. 5A**). As far as the time series of docking results differing in length among the cosmonauts, the slope of the reliability score was estimated individually and averaged afterwards (**Fig. 5B**). No general change in reliability over mission time was found.

DISCUSSION

The important finding of our study was to confirm a sufficiently high level of the cosmonauts' skill to manually dock a Soyuz or Progress on the ISS. Whereas once a serious problem occurred on Mir, the actual skill maintenance program on ISS is sufficient. In this respect the evaluation of cosmonauts' expectable operational reliability was of main interest for agencies and the responsible space operations staff. In the Russian space medical institute IBMP (Institute of Biomedical Problems) in Moscow, scientists have investigated that topic since the early sixties. Under the lead of Vyacheslav I. Myasnikov, Vyacheslav P. Salnitski, Albert P. Nechaev, and others, research focused on the mission-relevant operation of manual docking of a spacecraft on a space station. Under the title "maintenance of cosmonauts'

workability", this research was continuously implemented in the Russian long-term space research program. The research did not ignore that automated docking is preferable, but emphasizes that when technology fails, manual docking may be required. Therefore, cosmonauts should be informed regarding their current skill level to successfully conduct manual docking. In a fruitful long-term cooperation with the German Aerospace Center, assessment of the psychophysiological effort during docking training was developed as a robust, practical, and valid component in the operator reliability model. Joined analyses with the performance data improved our understanding regarding interactions between effort and performance. Moreover, we derived theoretical hypotheses and models. Given the specific circumstances of space research, particularly the low number of test subjects traveling to space, a very long time was needed to collect statistically valuable data. Indeed, this manuscript presents data, analyses, and conclusions spanning two and a half decades of docking research in space. As the human performance analyses were presented in a former publication, we focused on assessing psychophysiological data and the integration of performance and effort data into reliability models.

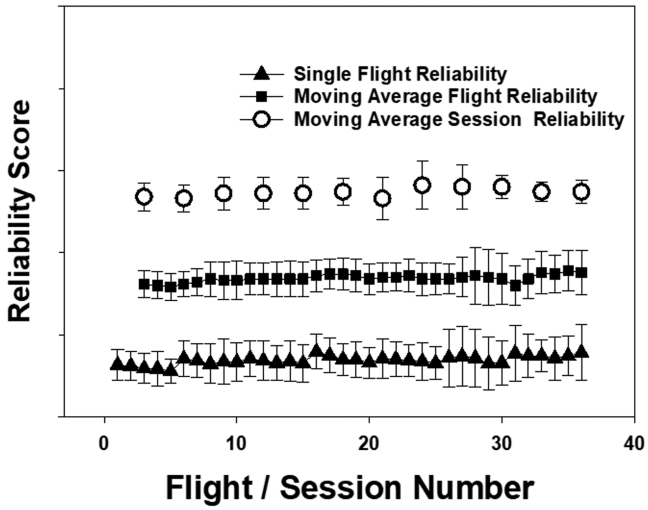

Fig. 4. Reliability scores over single training tasks, averaged training tasks, and averaged over training sessions. Session numbers are flight numbers divided by three. The y-axis scales for the three time series are changed to keep their visibility in the figure; they have the same mean score values.

Heart rate was permanently lowered under space conditions following the acute adaptation to weightlessness. We propose that this finding, which coincides with our own former results and changes in hemodynamics and volume status,[2] is not a sign of reduced workload. HR is a frequency-domain measure and negatively correlated to HPD. HPD, as a time-domain measure, covariates directly with the other effort indicators. Skin conductance appeared to react absolutely differently. From ARP research we know that not all people react with the SCL, but people of one specific ARP class react predominantly with SCL and FT. This is a strong sympathetic reaction with a vascular component. As the cosmonauts mostly demonstrated an autonomic stable pattern in all mission phases and the vascular responder[12] did not occur in the actual cohort (not presented here), the increased SCL may indicate a slightly higher periphery sympathetic level in flight. However, integrating the single psychophysiological parameters into the PAV takes into consideration the individual ARP and nullifies the single SCL response. Also, the inverted relationship of HR and HPD is taken into account.

The relation between docking accuracy and arousal indicated by HPD changes in flight is shown in Fig. 2. The "working range" seems to be reduced in space compared to terrestrial measures together with substantial shape changes of the three-dimensional inverse U-relationship. Comparing different single parameters as effort indicators suggests that under identical environmental conditions their time course is similar, albeit at different numerical levels. However, when the environment changes toward space conditions, effort parameters exhibit a differential response. On Earth, decreasing heart rate is usually associated with increasing pulse transit time. We observed the

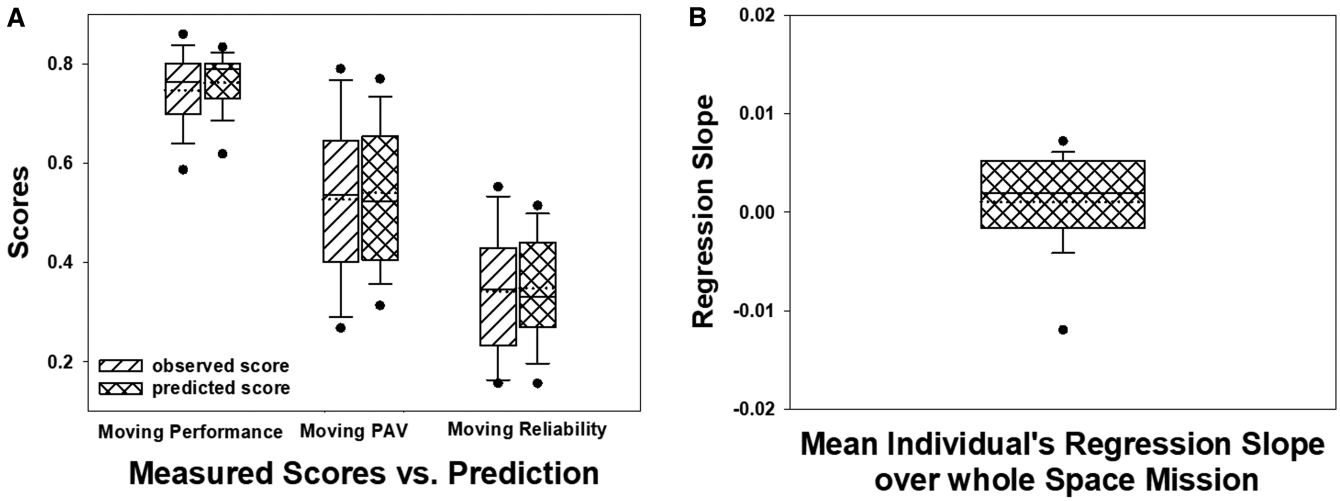

Fig. 5. Expected reliability during next training: A) linear regression prediction of performance, load, and reliability scores; and B) mean slope of reliability scores over all the missions, averaged over all astronauts, not significantly different from zero.

opposite in space. In summary, space-specific effects on physiological readouts have to be taken into account when using such parameters as workload indices.

Proceeding to more integrated, model-based indicators of workload, we applied the PAV. The model, which was developed and tested under terrestrial conditions in two larger cohorts,[12] was intended to better compare effort between cosmonauts. However, empirical evidence from space is limited. While 32 persons are a considerably large sample in space research, terrestrial studies sometimes include more than 1000 participants. Since statistical confirmation is still impossible, we have to rely on plausibility inspections. It is plausible that the workload level in space is increased, as indicated by the PAV. It is also plausible that in space, PAV follows different loads of the ARP assessment procedure.

Further analyses and modeling of reliability indicators shown here are based on PAV. A main goal of assessing docking training reliability is to predict successful real docking maneuvers. To leave the important operation undisturbed, real docking maneuvers were not routinely monitored psychophysiologically. As an exception, cosmonauts' speech comments during the docking were content analyzed and partially frequency analyzed during a period in the 1990s. Since all voice frequencies remained in the "green range", monitoring was discontinued.

The project "PILOT" on board the space station assessed psychophysiological parameters during docking training. An important finding[21] was that onboard training is absolutely necessary because manual docking skills deteriorate within 3 mo without practice. A series of five training flights focusing on the most difficult part, the docking final approach and contact, was recommended for at least once a month. The importance of having the skill to control an object with 6 degrees of freedom was emphasized. This skill, which is never applied on Earth and can only be trained on simulators, quickly fades without practice. There is a need of repeated pattern drill, as for playing a piano. Because in the last seconds docking control mistakes can lead to a catastrophe, skill automation is crucial. The psychological problem of such training is that one is training a skill that might never be needed in reality. Motivation is a central aspect. Situational variations may lead to fluctuating training results. For a prediction, we recommend extracting the best training results (performance) during a training session and analyzing the required effort, which in Sal'nitskiĭ's terminology is called "psychophysiological costs". Whereas a moving window over single tasks without selection still provides a stochastic curve, the session scores, selecting the best session results, splined over a moving window, seem to be a robust, easy, and practical approach. As Figs. 5A and B illustrate, clear tendencies in the reliability development over time can be recognized. From a practical point of view, using these splining methods, one will receive a more valid and reliable objective evaluation of each cosmonaut's docking performance after each training session, taking into consideration the trainings process before.

Using a simple statistical prediction approach based on regression analysis, we showed that reliability as well as individual performance and load scores did not deteriorate during the space missions (Fig. A; online, https://doi.org/10.3357/AMHP.5745sd.2021). We confirmed the finding for each cosmonaut individually (Fig. A). A beta level of 0.2, usually accepted as a verified rejection of an H_0, was not reached, indicating rather an increase in reliability. Thus, we assume that sufficient reliability levels were attained before spaceflight.

From a practical point of view, this may be the most important results of our analysis. Sal'nitskiĭ's model for the evaluation of an operator's reliability was further developed and tested, which turned out to be sensitive as well as robust enough for a practical application in this critical operational task, the manual docking maneuver.

ACKNOWLEDGMENTS

First of all, we thank all the cosmonauts who participated in the experiment. We thank in detail cosmonaut Prof. Dr. Reinhold Ewald for his professional support in evaluating the importance and difficulty of the manual docking maneuver. We are also thankful to the German Aerospace Center (DLR) for the continuous support of the project (to the first author: DARA–Grants 50WB9128, 50WB93401, 50WB93401-ZA; DLR–Grant 50WB 96220; and KIE–50WP0306, 50WP0501, 50WP0602, 50WP0603, 50WP1104, 50WP1304, 50WP1609). Greatest thanks go in memory to Vyatcheslav P. Salnitski (deceased 2016), who pioneered the complex IBMP-research on docking training and educated the first author.

Financial Disclosure Statement: The authors have no competing interests to declare.

Authors and Affiliations: Bernd Johannes, Dr. rer. nat., Sarah Piechowski, M.Sc., Hans-Juergen Hoermann, Dr. phil., and Jens Jordan, Prof. Dr. med., Head, Institute of Aerospace Medicine, German Aerospace Center (DLR), Cologne, Germany; Sergey V. Bronnikov, Dr. eng., S. P. Korolev Rocket and Space Corporation "Energia", Moscow, Russia; Juri A. Bubeev, Prof. Dr. med., Tatyana I. Kotrovskaya, Dr. med., and Daria V. Shastlivtseva, Dr. eng., Institute for Biomedical Problems (IBMP) of the Russian Academy of Sciences, Russian Federation State Research Center, Moscow, Russia; and Jens Jordan, Head of Aerospace Medicine, University of Cologne, Cologne, Germany.

REFERENCES

1. Albert J. Bayesian computation with R. New York: Springer-Verlag; 2009.
2. Baevsky RM, Baranov VM, Funtova II, Diedrich A, Pashenko AV, et al. Autonomic cardiovascular and respiratory control during prolonged spaceflights aboard the International Space Station. J Appl Physiol. 2007; 103(1):156–161.
3. Bell J, Holroyd J. Review of human reliability assessment methods. Darbyshire (UK): Health and Safety Laboratory; Health and Safety Executive; 2009.
4. Bosch Bruguera M, Fink A, Schröder V, Dessy E, van den Berg FP, et al. Assessment of the effects of isolation, confinement and hypoxia on spaceflight piloting performance for future space missions - The SIMSKILL Experiment in Antarctica. Acta Astronaut. 2021; 179:471–483.
5. Bubeev YA, Bronnikov SV, Kotrovskaya TI, Dudukin AV, Schastlivtseva DV, et al. First results of the experiment Pilot-T onboard ISS. XVI Conference on Space Biology and Medicine with International Participation, Young Scientists School; 5–8 December 2016; Moscow. Moscow (Russia): Institute of Biomedical Problems RAS; 2016.
6. Bubeev YA, Usov VM, Sergeev SF, Kryutshkov BI, Mihailyuk MB, Johannes B. Results of space experiment PILOT-T simulating the human-robot interactions on the lunar surface. Aerospace and Environmental Medicine. 2019; 53(7):65–75 .

7. Ellis SR . Collision in space. Ergon Des. 2000; 8(1):4–9.
8. ESA. Lebenslauf von Thomas Reiter. 2020. [Accessed 05.06.2020]. Available from https://www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/Astrolab_German/Lebenslauf_von_Thomas_Reiter.
9. Gaillard AW . Comparing the concepts of mental load and stress. Ergonomics. 1993; 36(9):991–1005.
10. Johannes B, Bronnikov S, Bubeev Y, Dudukin A, Hoermann H-J, et al. A tool to facilitate learning in a complex manual control task. Int J Appl Psychol. 2017; 7(4):79–85.
11. Johannes B, Bronnikov SV, Bubeev YA, Kotrovskaya TI, Shastlivtseva DV, et al. Operational and experimental tasks, performance, and voice in space. Aerosp Med Hum Perform. 2019; 90(7):624–631.
12. Johannes B, Gaillard A . A methodology to compensate for individual differences in psychophysiological assessment. Biol Psychol. 2014; 96:77–85.
13. Johannes B, Rothe S, Gens A, Westphal S, Mulder E, et al. Psychophysiological assessment in pilots performing challenging simulated and real flight maneuvers. Aerosp Med Hum Perform. 2017; 88(9):834–840.
14. Johannes B, Salnitski VP, Dudukin AV, Shevchenko LG, Shebuchev AE, Bronnikov SV . Performance assessment in the experiment PILOT onboard space stations Mir and ISS. Aerosp Med Hum Perform. 2016; 87(6):534–544.
15. Kim JW, Jung W . A taxonomy of performance influencing factors for human reliability analysis of emergency tasks. J Loss Prev Process Ind. 2003; 16(6):479–495.
16. Lager C . Pilot reliability. Stockholm: PECAB; 1974.
17. Park J, Jung W, Ha J, Shin Y. Analysis of operator's performance under emergencies using a training simulator of the nuclear power plant. Reliab Eng Syst Saf. 2004; 83(2):179–186.
18. Peake T. Ask an astronaut. London (UK): Arrow Books; 2018.
19. Richter P, Wagner T, Heger R, Weise G. Psychophysiological analysis of mental load during driving on rural roads – quasi-experimental field study. Ergonomics. 1998; 41(5):593–609.
20. Sal'nitskiĭ VP, Bobrov A, Dudukin AV, Johannes B . Reanalysis of operators' reliability in professional skills under simulated and real space flight conditions, Proceedings of the 55th IAC Congress, 4–8 Oct. 2004; Vancouver, Canada. Paris (France): International Astronautical Federation; 2004.
21. Sal'nitskiĭ VP, Dudukin AV, Johannes B . Evaluation of operator's reliability in long-term isolation (The "Pilot" Test). In: Baranov VM , editor. Simulation of extended isolation: advances and problems. Moscow: Slovo; 2001:30–50.
22. Sal'nitskiĭ VP, Myasnikov VI, Bobrov AF, Shevchenko LG . Integrated evaluation and prognosis of cosmonaut's professional reliability during space flight. Aviakosm Ekolog Med. 1999; 33(5):16–22.
23. Swain AD . Human reliability analysis: need, status, trends and limitations. Reliab Eng Syst Saf. 1990; 29(3):301–313.
24. Tim Peake launch as it happened. The Guardian; 15 Dec. 2015. [Accessed 31. 05.2020]. Available from https://www.theguardian.com/science/across-the-universe/live/2015/dec/15/tim-peake-launches-into-space-live.
25. Wickens CD . Multiple resources and mental workload. Hum Factors. 2008; 50(3):449–455.
26. Williams CE, Stevens KN . On determining the emotional state of pilots during flight: an exploratory study. Aerosp Med. 1969; 40(12): 1369–1372.
27. Wilson GF . An analysis of mental workload in pilots during flight using multiple psychophysiological measures. Int J Aviat Psychol. 2002; 12(1):3–18.