Performance Assessment in the PILOT Experiment On Board Space Stations Mir and ISS

Bernd Johannes; Vyacheslav Salnitski; Alexander Dudukin; Lev Shevchenko; Sergey Bronnikov

BACKGROUND: The aim of this investigation into the performance and reliability of Russian cosmonauts in hand-controlled docking of a spacecraft on a space station (experiment PILOT) was to enhance overall mission safety and crew training efficiency. The preliminary findings on the Mir space station suggested that a break in docking training of about 90 d significantly degraded performance.

METHODS: Intensified experiment schedules on the International Space Station (ISS) have allowed for a monthly experiment using an on-board simulator. Therefore, instead of just three training tasks as on Mir, five training flights per session have been implemented on the ISS. This experiment was run in parallel but independently of the operational docking training the cosmonauts receive.

RESULTS: First, performance was compared between the experiments on the two space stations by nonparametric testing. Performance differed significantly between space stations preflight, in flight, and postflight. Second, performance was analyzed by modeling the linear mixed effects of all variances (LME). The fixed factors space station, mission phases, training task numbers, and their interaction were analyzed. Cosmonauts were designated as a random factor. All fixed factors were found to be significant and the interaction between stations and mission phase was also significant.

DISCUSSION: In summary, performance on the ISS was shown to be significantly improved, thus enhancing mission safety. Additional approaches to docking performance assessment and prognosis are presented and discussed.

KEYWORDS: spaceflight, manual docking training, Soyuz, Progress, individual styles, prediction.

Johannes B, Salnitski V, Dudukin A, Shevchenko L, Bronnikov S. *Performance assessment in the PILOT experiment on board space stations Mir and ISS*. Aerosp Med Hum Perform. 2016; 87(6):534–544.

Humans under the conditions of long-term spaceflight are exposed to numerous stress factors, e.g., environmental-physical, social, and informational. These factors are considered to represent a main risk for failures and errors within the complex crew-spacecraft system.[10,11] Preliminary findings on the Mir space station suggested that a break in docking training of about 90 d significantly decreased performance.[13] Therefore, the assessment of cosmonaut's performance and reliability of docking skills is considered to be an important way to analyze the crew's operational reliability.[9] In the present study, we focused on the manual docking maneuver. A cosmonaut's reliability in this mission-relevant operation has central importance for the operational reliability of the whole man-machine system.

In the seventies, Komotski and colleagues started a scientific program for objective performance assessment during crew activities, among them docking training.[7,8] This work was then continued with an IBMP-RSC Energia-DLR collaborative project: the space experiment PILOT. The aim was to develop a PC-based autonomous research docking simulator and to investigate different approaches to evaluate an operator's reliability in manual docking.[13,17] This methodology was applied and tested in the PILOT experiment on the Mir space station, the International Space Station (ISS), and in several terrestrial ground-based experiments in space analogues (e.g., isolation, bedrest, immersion, etc.). The expert knowledge-based

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

This manuscript was received for review in July 2015. It was accepted for publication in February 2016.

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

DOI: 10.3357/AMHP.4433.2016

coefficient of exactness (Kt) was implemented into the regular docking training of cosmonauts as well as into the software of the PILOT experiment and thereby became the "gold standard" for performance evaluation of this maneuver. We retain the name of the index as "Kt" because it has fundamental relevance in Russian performance evaluation in all publications. The "K" stands for coefficient and the "t" for exactness (Russian: tochnost). To validate the Kt, several statistical methods were implemented. These methods should integrate the numerous raw parameters into one objective "quality" indicator based on data and not assumptions. Canonical correlation analyses for the comparison of physiological data[14–16] were tested as well as exploratory factor analyses for the separate evaluation of the performance data and the psychophysiological load.[5,6] Confirmatory factor analyses were then performed for the verification of the latter. The main approaches and methods used for the assessment of performance are described in this paper. The results presented herein are based on data obtained during spaceflight experiments on both Mir and the ISS.

METHODS

The performance evaluation of a spaceflight maneuver was originally prepared by Salnitski and colleagues for the situation of a manually controlled redocking flight. This maneuver becomes necessary if the docking point on the space station (SS) used for automated docking is blocked by a spacecraft (SC), but will be required for another approaching SC. This redocking flight can start and end at several existing docking points of a SS. The SS has had several changes in its configuration during its life cycle. Therefore, an automated program for each flight path is difficult to maintain. Manual control of redocking flights was the routine procedure during the Mir period and continues still on the ISS. Training and skill maintenance of manual control and docking of a SC on a SS has always been a fundamental part of Russian cosmonauts' education. During the Mir period, research simulator software was developed by the working group of Salnitski et al. in the IBMP (mainly by Jury Shlykov). For the ISS epoch, the research simulator software was provided by RSC Energia and was also used for the regular docking training of cosmonauts.

The standard position of a SC is to be docked at the SS. A standard redocking flight is divided into five flight phases. The "flight-off" (flight phase 1) begins with the moment of decoupling of the SC from the SS and ends when the SC has reached a safe distance from the SS (30–40 m). The "stabilization-1" phase (flight phase 2) occurs when the SC is within the safety distance and is correctly orientated toward the SS prior to the "flight-around" (flight phase 3). The "flight-around" phase starts when the SC leaves the "stabilization-1" position and ends at a second "stabilization-2" position. During the "flight-around," the distance to the SS has to be kept within an optimal and safe corridor. The SC has to be kept continuously oriented perpendicular to the body of the SS. The required sideways flight with the SC is one of the most difficult maneuvers of the redocking flight. Any collision with parts of the SS has to be avoided and, with respect to the actual configuration of the SS, the requested flight path and the safe distance differ. The "stabilization-2" phase (flight phase 4) prepares the SC for the final docking approach. The SC has to be stabilized at the center line of the docking point while at safety distance. The orientation of the SC can be best prepared at this distance (lowest angle errors). The "final approach" (flight phase 5) begins when the SC leaves the "stabilization-2" position and ends with the moment of contact with the SS, the "docking." The "docking" phase is not considered to be a flight phase and is therefore evaluated separately. It is, however, the most important and critical moment of the redocking flight. The evaluation score for the fifth flight phase ("final approach") was in practice often considered the most important as it summarizes the final approach and moment of contact. Therefore, our analyses focus on this indicator (Kt_5, described in detail in **Appendix A**, which is available online; 10.3357/amhp.4433sd.2016).

In the evaluation of redocking flight quality, 12 parameters (**Table I**) play a central role. These are simply a set of 12 physical-mathematical parameters that describe the position and the motion of the SC and SS with regard to each other. The nomenclature is illustrated in **Fig. 1**.

The main measurements for the contact moment are the distances in y- and z-axes, the relative speeds along all axes, and the angles between axes of the docking compartments of the spacecraft and the space station. For the flight-around phase, the most relevant parameters are the optimal distance from the space station and the continuously optimal orientation of the spacecraft toward the space station. During the final approach phase, the following parameters are analyzed: deviations from the center line, optimized speed toward the station with respect to the actual distance.

The Kt represents an expert knowledge-based common evaluation of a complete redocking training flight. The mathematical apparatus was published in parts by Dudukin et al.[4] and is presented in detail in our Appendix A online (10.3357/amhp.4433.2016). The general idea is that safety ranges were

Table I. Description of Performance (Raw Data).

POSITION	FIRST DERIVATION: MOTION	DESCRIPTION
ρ	$d\rho/dt$	Distance between the visor of the SC and the docking point of the of the SS/approach speed
φ_1	$d\varphi_1/dt$;	Yaw, course angle (y axis, Y_1) of the SC with regard to the SS
θ_1	$d\theta_1/dt$	Pitch angle (z axis, Z_1) of the SC with regard to the SS
φ_2	$d\varphi_2/dt$;	Yaw, course angle (y axis, Y_2) of the SS with regard to the SC
θ_2	$d\theta_2/dt$	Pitch angle (z axis, Z_2) of the SS with regard to the SC
γ	$d\gamma/dt$	Bank angle (x axis, $X_1 = X_2$) between SC and SS

An index of 1 is related to the space craft, an index of 2 is related to the space station. SC: space craft; SS: space station.

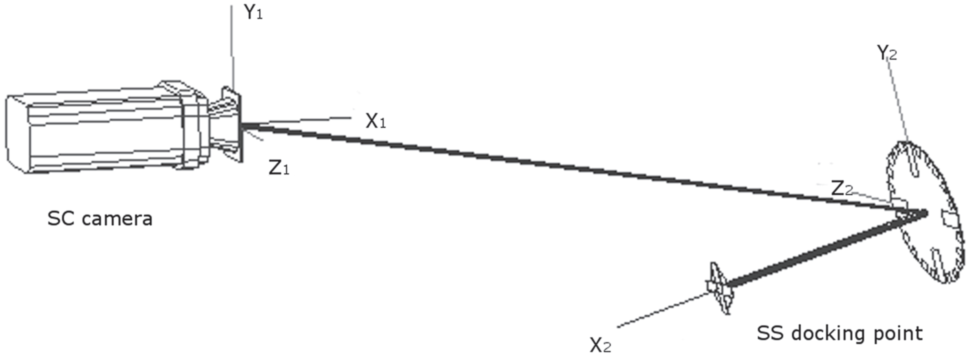

Fig. 1. Definition of coordinate system for the estimation of the relative movement parameters between the space craft (SC) and the space station (SS) docking target. In the figure the position of the SC's telecamera is given; however, the parameters are calculated with regard to the docking apparatus of the SC.

defined for all controllable parameters and any deviation from the range was registered per time interval (safety ranges are given in Appendix A, Table AI online; 10.3357/amhp.4433sd.2016). **Fig. 2** illustrates the safety range for an example flight track around the Mir space station.

For each ith flight phase a quality coefficient Kt_i was calculated. These coefficients were combined to give a weighted average as common Kt (Eq. 1).

$$Kt = \frac{\sum_{i=1}^{m} \beta_i \left(1 - e^{t_i/t_0}\right) Kt_i}{\sum_{i=1}^{m} \beta_i \left(1 - e^{t_i/t_0}\right)} \qquad \text{Eq. 1,}$$

where m = number of flight phases (for the complete redocking flight m = 5), $\beta_1 = [1, 1, 2, 3, 3]$, $t_0 = -5$, and t_1 = duration of ith flight phase. The Kt represents an expert knowledge-based common evaluation of a complete redocking training flight. However, it is applicable also for the shorter training flights in the experiment, consisting only of phases 3 to 5. Kt_5 is the Kt_i with i = 5.

In the first statistical approach for integrating several raw parameters of a docking flight into one coefficient for "work quality," canonical correlation analysis was used,[14–16] but will not be described herein again. In a second statistical approach, exploratory factor analyses (FA) were used. The aim was not to find "common factors" behind the raw data, but rather to create an orthogonal reference frame to allow for an orthogonal vector sum integration of the factor scores.[12, p.482] For each training flight, the experimental simulator software provided the raw parameters given in Table I for each flight phase. Additional parameters of fuel consumption and the evaluation of the optimal use of fuel were also given. All approaches, including fuel parameters, were excluded from the herein presented performance analyses.

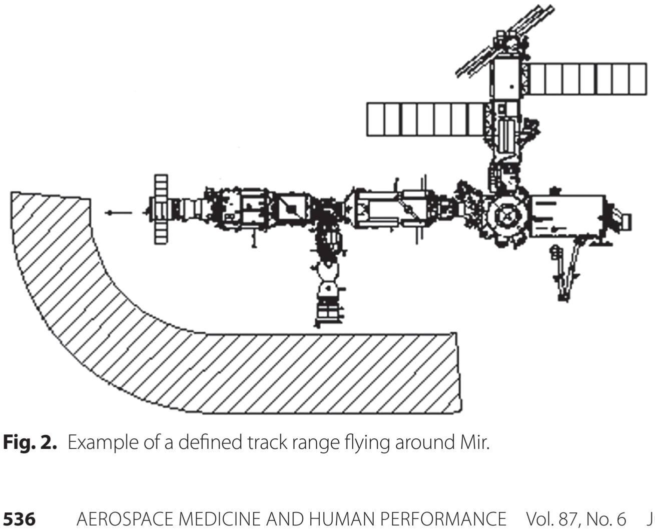

Fig. 2. Example of a defined track range flying around Mir.

A FA provides a reference frame (usually an orthogonal dimensional space) that explains the most variance of the numerous raw parameters with a reduced set of factors (dimensions). The herein used approach accepted as factors all eigenvectors with substantial variance, not only those with an eigenvalue larger than 1 (Kaiser-Guttman-Rule[12, p.482]). The full-factor solution (all eigenvectors are "factors") is accepted as the only explanation of the overall variance of the data[12, p.465] and this approach needs as much as possible explained variance. The presented approach included all eigenvectors, which explain together 90% of the cumulative variance.

The flight around the space station or all other possible approach flights toward the area of stabilization (stabilization2) prior to the final approach differ for all training situations. The stabilization2 phase is the first standardized and ultimate flight phase for all docking tasks. Therefore, exploratory FAs were run for the last three flight phases separately (for detail see **Appendix B** online; 10.3357/amhp4433sd.2016). Afterwards a set of multiple regression functions was calculated separately for each performance factor and each flight phase to allow for future training flight evaluations (given in SPSS script style in Appendix B online; 10.3357/amhp4433sd.2016).

Confirmatory FA (AMOS 7.0, SPSS, IBM) was used to investigate whether the factor-analytic performance model has the same general structure for all cosmonauts. Furthermore, the model was individualized for each cosmonaut, differentiating between the docking skills of the cosmonauts. In contrast to the exploratory FA, which looks for factors in a particular data set, the confirmatory FA assumes the existence of a given factor structure and tests how the raw data fit this factor model. For the confirmatory FA, the raw data are required as an input to test whether the constructed vector space is reliable and stable across different data samples. A model of confirmatory FA represents a set of linear equations also known as a "structural equation model." However, AMOS provides a graphic user interface for modeling the equation systems of the confirmatory FA, resulting in graphs. The confirmatory FA models were developed in an iterative process and were then applied separately to each cosmonaut. The Chi-squared test was used to examine the fit of the models.

The cosmonauts are assumed to have individual styles of control during docking maneuvers. This could be assessed by means of the different confirmatory FA models and the fits for each cosmonaut. Another approach to assess individuality groups is the use of cluster analyses. The WARD method was used, which is known to detect very robustly the number of clusters in a certain data mass.[3] The pairs of single data sets are analyzed and a measure of distance is calculated based on Euclid squares of differences in the single parameters. Groups of data sets with low distances are assigned to one and the same cluster, herein a group of a certain control style.

In most fields of science, the *P*-value hypothesis test has established a monopoly on statistical reporting. An alternative measure is conveyed by a Bayesian hypothesis test, which prefers the model with the highest average likelihood.[1,2] Bayesian multilevel modeling provides probabilities for expected next events (more in detail in **Appendix C** online; 10.3357/amhp.4433sd.2016). This could be of great importance for the prediction of the next future performance of a cosmonaut.

The following assumptions were aimed to be tested by developing and testing different performance evaluation methods:

- We assumed an increased performance level on ISS.
- The different integration approaches should provide correlating indicators, however, assessing different aspects of performance.
- New performance evaluation summarizing over whole mission phases can be provided.
- Individual work styles can be assessed.
- Statistical predictions of expectable performance can be provided.

The PILOT experiment, part of the Russian long-term space research program, was jointly developed between scientists and engineers of the IBMP, RSC Energia, and DLR. The IBMP developed the initial scientific idea and the first research simulator software. IBMP was the general lead for the development of performance evaluation methods. RSC Energia provided the onboard computer, hand controls, onboard integration, the space transportation, and crew time on board. Since the beginning of the ISS epoch, RSC Energia has provided the high-quality simulation software. The DLR provided the psychophysiological assessment systems and methods, and supported the data analysis. The "PILOT" experiment was approved both by the local IRB (IBMP) and the Human Research Multilateral Review Board (for ISS experiments).

Subjects

Russian male adults participated in the study. For the in-flight studies, 5 cosmonauts served as subjects on the Mir station, and 12 cosmonauts served on ISS.

Procedure

From 1996 to 2001 on the Mir station and 2008–2011 on the ISS, all Russian cosmonauts underwent three preflight (−1 mo, −10 d, −3 d prior to launch) and three postflight (+3 d, +10 d, +2 to 3 mo post-landing) experiments. The individual flight duration differed, but was around 6 mo (min 164, max 195 d). In flight, the cosmonauts executed the experiment on Mir sporadically, but on the ISS at regular monthly intervals.

The PILOT experiment aimed to investigate cosmonaut's skill in and performance of manual docking of a Soyuz spacecraft on the space stations (Mir and ISS) during different stages of long-term spaceflights. The experimental docking simulator challenged the cosmonauts with a series of docking flight tasks. For the dynamic and informational equivalence to real docking maneuvers, the simulation was based on mathematical models for real hand control of the Soyuz SC. The cosmonaut saw a synthesized view of the actual space station on the screen identical to the optical camera view of the real docking system. The required technical information was provided by RSC Energia and the experimental simulator was verified by RSC Energia with support from Russian cosmonauts. The quality of the computer model increased from the Mir period to the ISS epoch on a photographic level; however, the dynamics of the controlled SC remained identical. Original standard control handles were used for the experiments.

During the experimental docking flights no instruments for flight parameters or information about relative speed or distance to the SS were presented to the cosmonauts. Instead, they had to fly strictly based on the visual information on the screen. During the Mir period (1996–2000), three tasks were given per training session, whereas in the ISS epoch (2008–2011), five tasks had to be fulfilled. All tasks were different but their order remained identical for each experimental session. The tasks focused on the moment of docking and started at the end of different flights-around toward different docking points.

As primary outcome measures of performance the Kt, as well as the phase specific coefficient of exactness (Kt_5, assessing exactness of the final approach and the docking contact), were used as provided by the simulator software. Additionally, a pass/fail criterion was estimated. A docking was considered to be successful if all final parameters of distances and speeds during the docking contact were within given safety ranges (Appendix A online; 10.3357/amhp4433sd.2016).

The main statistical work was done with the SPSS for Windows package. The results presented herein were calculated using version SPSS v.20. For the comparison of performance level between Mir and ISS, nonparametric tests were run and linear mixed effect (LME) models were tested to confirm these results. Because the Kt and Kt_5 data were not normally distributed, it was deemed necessary to perform a Box-Cox transformation of these data. A Box-Cox transformation optimizes the exponent λ of an exponential transformation with the aim to result in a normal distribution of transformed data. It was then necessary to perform Box-Cox transformations of the Kt and Kt_5 data. The LME models included as fixed effects the stations, mission phases, and the flight number within a training session. The cosmonaut ID was set as random effect. Variances were allowed to differ among cosmonauts, and LME models were optimized according to the Akaike information criterion. A model was accepted if the residuals were not rejected as being normally distributed.

The comparison of the different approaches of performance assessment presented herein was performed by correlation analyses. Cluster analyses were used to detect particularities of individual cosmonauts in their docking skills. Bayesian analyses were carried out in the "R" statistical environment (version 2.9.2, www.r-project.org). The level for statistical significance was set to $\alpha = 0.05$.

RESULTS

The mean coefficient Kt was 0.634 (SD = 0.15) on the Mir station and 0.875 (SD = 0.09) on the ISS; the mean Kt_5 was 0.814 (SD = 0.23) on Mir and 0.839 (SD = 0.15) on the ISS. The Kolmogorov-Smirnov test rejected normality of the distribution of Kt and Kt_5 for both stations. The left-skewness indicated a dominance of higher performance values.

In the first step, nonparametric testing (Mann-Whitney U) was employed for a statistical comparison between stations. The Kt score was significantly different ($P < 0.001$), but not the Kt_5 ($P = 0.410$). However, in testing with the two-sample Kolmogorov-Smirnov test, both coefficients differed between the stations (Kt: $P < 0.001$; Kt_5: $P = 0.011$). **Fig. 3** presents the common performance score Kt of both stations over the mission phases (preflight, in-flight, postflight) and over a training session (in-flight data only).

For all further statistical testing, the result of the Mann-Whitney test will be given; however, for comparisons between mission phases and between the flight tasks of a training session, a LME model is the appropriate and desired analysis. The residuals of LMEs with the original Kt and Kt_5 data were not normally distributed and, therefore, the Kt and Kt_5 were Box-Cox transformed. The Kt scores could be transformed into a value Kt_t = $(Kt - 0.19 + 1)^{4.35}$, which was not rejected and was normally distributed ($P = 0.088$). The result of the LME with the transformed values determined that the residuals were normally distributed ($P = 0.108$). The fixed effects of station [df: num 1, denum: 16,925, $F(1, 16{,}925) = 75.614$, $P < 0.001$], mission phase [df: num 2, denum: 693,830, $F(2, 693{,}830) = 8.949$, $P < 0.001$], and flight number [df: num 4, denum: 682,927, $F(4, 682{,}927) = 83{,}514$, $P < 0.001$] were significant and the interaction between station and mission phase was also significant [df: num 2, denum: 693,901, $F(2, 693{,}901) = 18.799$, $P < 0.001$]. However, no significance occurred for the interaction of mission phase and flight number [df: num 8, denum: 683,008, $F(8, 683{,}008) = 0.984$, $P < 0.504$]. The Akaike's Information Criterion (AIC) was about 2782.48. Excluding the insignificant interaction of mission phase and flight number provided an AIC of 2790.57, indicating the model was slightly worse. In summary, performance was different both between the stations and between mission phases. Additionally, the performance changes between mission phases were different for both stations. The task performance between the different tasks within a training session differed. These differences remained constant over the mission phases.

For the Kt_5 no successful box transformation for normalization was found. The residuals of any applied LME were never normally distributed. However, after exclusion of outlier values of Kt_5 (occurring only in the Mir data; $|Kt_5| > > 3*SD_{Kt}$, remaining $n_{MIR} = 92$, $n_{ISS} = 610$) a LME model was found with normally distributed residuals ($P = 0.113$). Although the effect of mission phase did not reach statistical significance ($P = 0.087$), the effect of flight number was confirmed to be significant ($P < 0.001$).The effect of station ($Kt_{5,MIR} = 0.853$, SD = 0.19; $Kt_{5,ISS} = 0.839$, SD = 0.15) was insignificant [$F(1, 16) =$

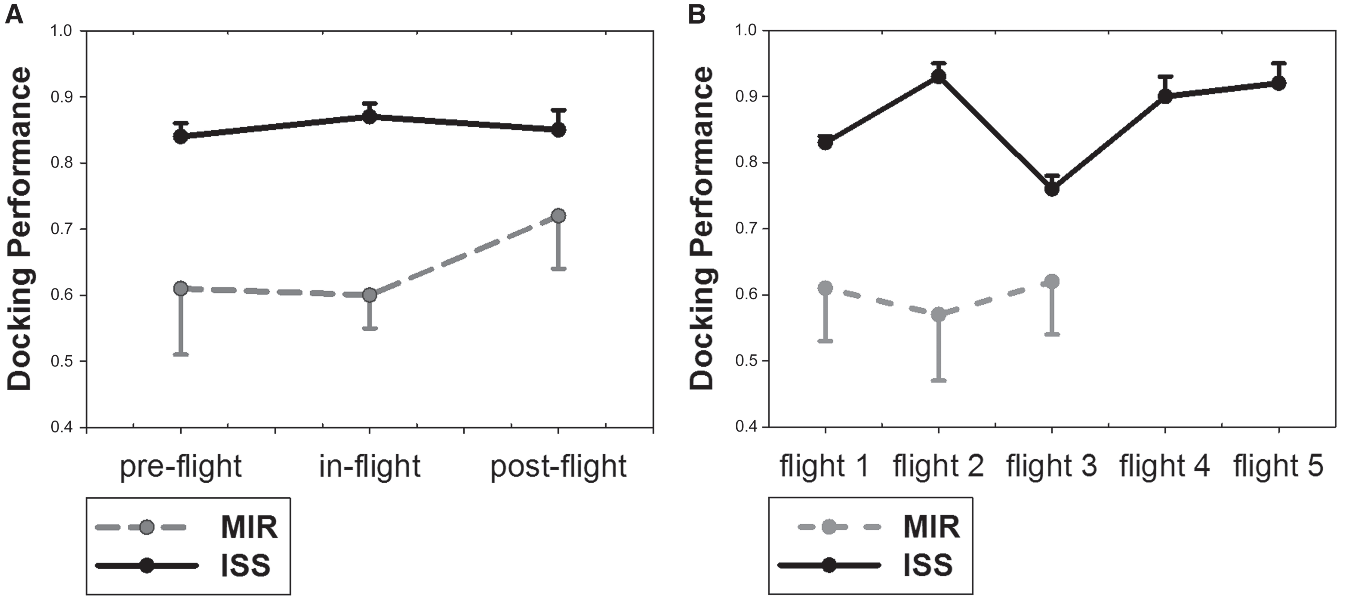

Fig. 3. Mean performance (Kt) on Mir and the ISS. Left: over mission phases; right: over flight tasks within a training session.

0.335, $P = 0.571$], but its interaction with mission phases was significant [$F(2, 677) = 9.895$, $P < 0.001$]. Although the flight number effect was significant [$F(4, 667) = 15.595$, $P < 0.001$], its interaction with the mission phase was not [$F(8, 667) = 0.307$, $P = 0.963$]. For the Kt_5 model, the AIC was 10,402.9, which was nearly four times larger than for the Kt models, thus indicating that the Kt_5-model was much less accurate than the Kt models.

The different flight phases were factor analyzed separately. Parameters were selected for the different flight phase that best described changes in those flight phases, as described in detail below. The data set, cleaned from outliers, was used for modeling the reference frames. Only data obtained during spaceflights were included.

For the most relevant flight phase (moment), the docking contact, nine raw parameters were analyzed. Based on the cumulatively explained variances (see Appendix B online; 10.3357/amhp4433sd.2016) an eight-factor model for the $Kt_{f\_contact}$ was accepted. There were 11 variables used for the final approach phase FA. These variables were all standard deviations of raw parameters. A six-factor model for the $Kt_{f\_final\_approach}$ was accepted. The stabilization2 phase prior to the final approach was factor-analytic analyzed using 12 variables. A five-factor model for the $Kt_{f\_stabilization2}$ was accepted.

In the simulator software, the docking contact performance is integrated together with the last few meters of the final approach into the Kt_5 coefficient. Therefore, an additional FA was run including the variables of the docking contact analysis and the final approach analysis. An eight-factor model was accepted for the coefficient Kt_f_Kt_5 as analogue of the original Kt_5 coefficient.

Analogous to the original Kt, the factor-analytic performance scores were averaged across phases to provide a common factor-analytic coefficient of exactness, as summarized in Eq. 2.

$$Kt_f = (Kt_{f\_stabilization2} + Kt_{f\_final\,approach} + Kt_{f\_contact})/3 \qquad \text{Eq. 2}$$

The factor-analytic common coefficient of exactness Kt_f was significantly higher on the ISS compared to the Mir station [$Kt_{t\_MIR} = 0.755$, $Kt_{f\_ISS} = 0.812$; Mann-Whitney U, $P < 0.001$; LME: $F(1, 12) = 16.68$, $P = 0.001$, normally distributed residuals].

Fig. 4 illustrates the differences of the flight phase wise factor-analytic coefficients of exactness. The $Kt_{f\_contact}$ for the docking contact moment was significantly increased on the ISS [Mann-Whitney U, $P < 0.001$; LME: $F(1, 18) = 23.66$, $P < 0.001$, normally distributed residuals]. The $Kt_{f\_stabilization2}$ for the stabilization phase was also significantly higher on the ISS [Mann-Whitney U, $P < 0.001$; LME: $F(1, 13) = 9.377$, $P = 0.008$, normally distributed residuals].

The residuals of the LME with the original $Kt_{f\_final\_approach}$ values did not distribute normally. After Box-Cox transformation of the original values, the residuals of the LME became normally distributed; however, the station effect was not significant [Mann-Whitney U, $P = 0.506$; LME: $F(1, 16) = 2.58$, $P = 0.128$]. The combination of the "docking contact" and the "final approach" into one FA provided a significant effect between the space stations and normally distributed residuals [$Kt_f\_Kt_{5,MIR} = 0.806$, $Kt_f\_Kt_{5,ISS} = 0.852$; Mann-Whitney U, $P = 0.011$; LME: $F(1, 15) = 20.95$, $P < 0.001$].

Table II presents the correlations among the different coefficients of performance. High correlation between factor-analytic and original expert coefficients can be considered as validation of the latter ones. Significant correlations were found for the coefficient of stabilization2 ($Kt_{f\_stabilization2}$) with the original phase 4 score Kt_4 and with both common coefficients (Kt and Kt_f). No correlation was found for the expert evaluation of the final approach phase (Kt_5) and its factor-analytic evaluation ($Kt_f\_Kt_5$).

Assuming that the standard Kt_5 coefficient combined the final approach and the docking contact moment, the confirmatory FA presented herein attempted to verify the $Kt_f\_Kt_5$ coefficient. In **Fig. 5** the four factor (ellipses) model is depicted. Of the 13 input variables, 11 (rectangles) of the exploratory FA were sufficient to explain the variance and to differentiate among subjects. Error terms (circles) completed the model. The different variants of this model were only allowed to have different interrelations among the four basic factors. In other words, the basic factors show significant correlation when in the model an interrelation arrow is present. In Model 41, illustrated in Fig. 5, all four basic factors (docking, final SC, pitch SS, yaw SS) were correlated.

In models 42 to 45, different interrelations of these basic factors were left out. No model was found to describe the individual data of any cosmonauts without any interrelation. In a former confirmatory FA approach (not illustrated here), the final approach and the docking contact moment were modeled separately. This former three factor model is assumed to be similar to the $Kt_{f\_final\_approach}$ evaluation and consists only of the lowest three factors of the given model. The different versions of the three-factor models are identified in **Table III** and **Table IV** with numbers in the 30s. The models sufficiently explain the variance of the obtained performance if P of the Chi-squared test is >0.2 and the model was assigned to fit the data for a certain cosmonaut.

For testing whether the different models are related to performance, the classic and newly developed performance indicators were compared between the fit and nonfit groups for all models. Table III presents the significances of the performance differences, illustrating that numerous models are related to the docking performance. For each cosmonaut the models were verified to fit or not (see Table IV). Excluding the cases where the number of available training flights was too small for any fit, one could recognize that the cosmonauts differed clearly in the fit of the models. This could be interpreted as differences in the personal styles of docking. Individual patterns of the hand control docking skill were also differentiated by cluster analysis (**Fig. 6**). The factor-analytic performance scores of the different flight phases were averaged for each cosmonaut. These averaged values were put into a WARD cluster analysis.

One large main group and four individual outliers could be identified. Strikingly, all outliers were cosmonauts from the

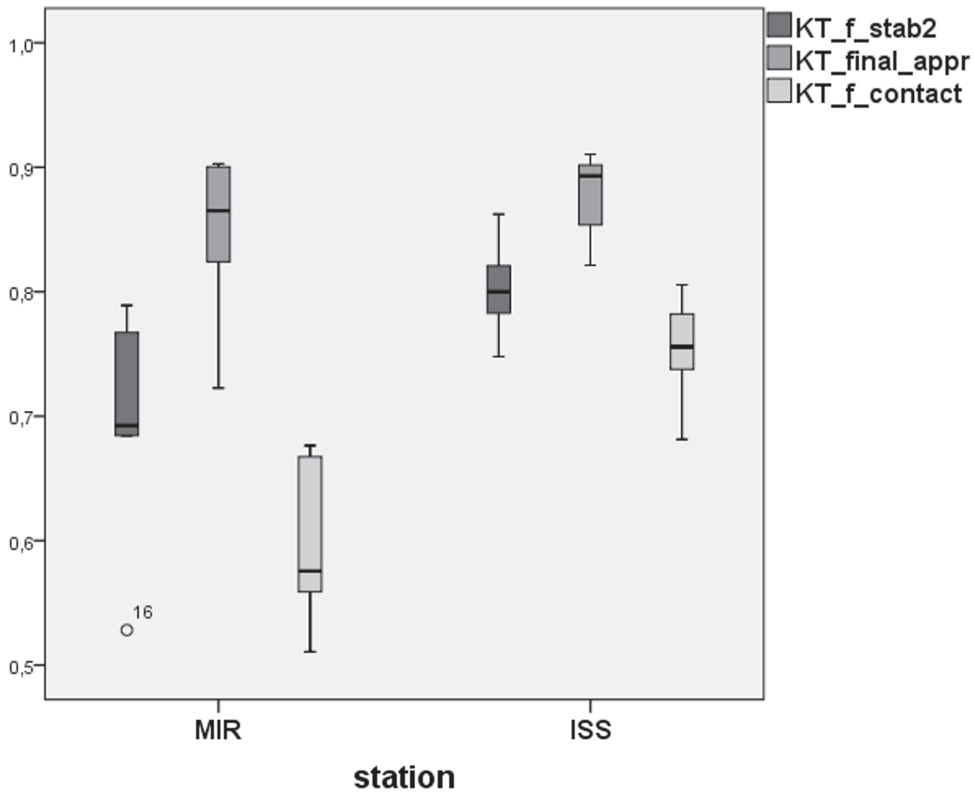

Fig. 4. Factor-analytic performance scores of three flight phases on Mir and the ISS. Light grey: $Kt_{f\_contact}$; medium grey: $Kt_{f\_final\_approach}$; and dark grey: $Kt_{f\_stabilization2}$.

Mir station (cc = 0.652, $P = 0.014$). The group of Mir cosmonauts was not only different from the group of ISS cosmonauts, but also clearly nonhomogeneous. The standard coefficients for performance Kt and Kt_5 were significantly different among the cluster groups (both: Mann-Whitney U, $P < 0.001$).

A main aim of all training and performance evaluation is the prediction of the expected performance of the next, usually the upcoming "real" docking. Bayesian statistics promises probability estimation for upcoming events. For this kind of analysis we used the pass/fail data. The cosmonauts' successes and failures with regard to some safety range criteria provided individual percentages of success. The mean percentage was significantly different between both stations (Mann-Whitney U, $P < 0.001$). Our Bayesian analysis starts with calculating a conditional probability as to whether the next (training) flight will be successful if the training flight before was successful. The expected docking success was found to be significantly higher on the ISS (Wilcoxon W = 9, $P < 0.03$). It is, however, necessary to mention that the level of expected success was still sufficient on the Mir station. **Fig. 7** illustrates in a graphic form the probability intervals of success in next docking maneuver for each cosmonaut. The pass-fail percentage represents the x-axis of this graph. The y-axis provides the expected probability for success and the respective range. The mean expected probability of

Table II. Pearson Correlations (r) and Significances (*P*) Between Expert Scores and Factor-Analytic Scores of Performance.

	Kt_f	$Kt_{f\_contact}$	Kt_{f_final}	Kt_5	Kt_{f_Kt5}	Kt_4	$Kt_{f\_stabilization2}$
Kt							
r	0.546	0.424	0.242	0.470	0.340	0.774	0.475
P	*	*	*	*	*	*	*
Kt_f							
r		0.683	0.612	0.104	0.765	0.447	0.603
P		*	*	0.061	*	*	*
$Kt_{f\_contact}$							
R			0.191	0.132	0.622	0.409	0.020
P			*	0.015	*	*	0.719
$Kt_{f\_final\_approach}$							
r				0.092	0.745	0.249	0.158
P				0.093	*	*	0.004
Kt_5							
R					0.064	0.271	0.097
P					*0.244*	*	0.078
Kt_{f_Kt5}							
R						0.295	0.149
P						*	0.007
Kt_4							
r							0.241
P							*

Correlation analysis among coefficients of exactness, * $P < 0.001$.

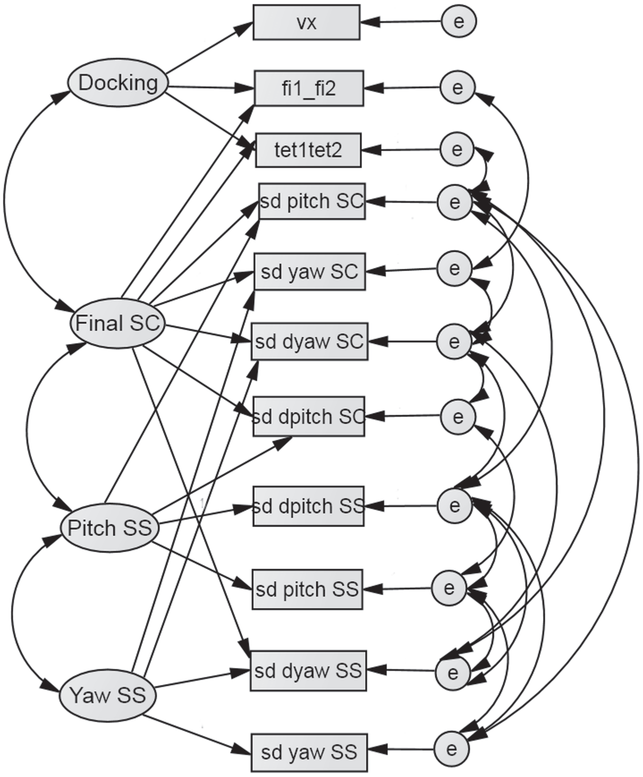

Fig. 5. The confirmatory factor model of spacecraft docking performance graphically represents a system of equations. A basic four-factor model (ellipses) is explained by 11 variables (rectangles) and the respective error terms (circles). A Chi-squared test confirms ($P > 0.2$) whether this model fits a data set or not.

successful docking for all cosmonauts was 80% (dashed horizontal line). However, the individual approach (dotted diagonal line) illustrates that a higher success probability is expected from cosmonauts with a higher training flight success.

DISCUSSION

Salnitski and his colleagues provided the very first computerized and autonomous onboard research simulator for an important and really complex space operation—the manual docking of a spacecraft on the Mir station.[13] Historically, this became necessary because the former training system was based on satellite connections and data transmission between the station and Earth. Therefore, this training system was not always available. A main result of this research demonstrates that performance level, assessed by means of the coefficient of exactness Kt, was, from a safety perspective, high enough on the Mir station. The greatest difficulties were found with the very first cosmonauts on the Mir station who were not sufficiently familiarized with the research simulator before the flight because the hardware arrived only during their spaceflight. However, one has to thank them because they made the research simulator run on board.

Preliminary results obtained on the Mir station[13] during some selected missions suggested that a break in training of about 90 d significantly degraded performance below the safety requirements. A comparison of the Mir period and the ISS epoch of the PILOT experiment demonstrated a significant improvement of experimental docking quality on the ISS (Table I, Fig. 2). A significant interaction (LME) between the stations and the flight phases underlines the more intensive preparation of the cosmonauts and their constantly high skills during the ISS epoch, whereas during the Mir period the cosmonaut's performance still increased after their flight, indicating a further training effect.

The work of Salnitski and colleagues with respect to the performance assessment was of striking importance. Thereafter, permanent new approaches were verified and compared with others for validation. Unfortunately, the capacity of data transfer between the station and Earth was limited during the Mir period and only condensed results were transferred. Therefore, the performance assessment was programmed to provide fixed results. The methodology used here is presented in detail for the first time and all post hoc analyses were oriented on validation of these results. Also, the raw data was successfully cross-validated due to inherent physical relationships. For example, a certain turn around the x-axis (bank) also increased the distance measures for the z-axis and so on. Integration of the mass of raw data, however, was based on assumptions and expert decisions. It remained an open question whether the definition of a certain safety range for a raw parameter was really optimal. Also, the integration of all single quality evaluations for single phases and then into a common parameter (Kt) was not based on data, but rather on the decision of the experts. The main advantage of this kind of performance evaluation was the fully mathematically described apparatus. The subjective evaluations of the instructors, based on their experience with the docking system, and the cosmonaut were of essential value,

Table III. Performance Differences for Confirmatory FA Model Fits vs. Nonfits.

MODEL	Kt_4	Kt_5	Kt	$Kt_{f\_contact}$	$Kt_{f\_final\_approach}$	$Kt_{f\_stabilization2}$	Kt_{f_Kt5}	Kt_f
31	0.147	0.002	0.849	0.233	*	0.001	0.008	*
32	0.024	0.094	0.466	0.343	0.003	0.560	0.036	0.017
33	*	0.131	0.013	0.011	0.541	0.660	0.085	0.073
41	0.602	0.234	0.444	0.033	0.089	0.019	0.532	0.503
42	0.091	0.000	0.001	0.019	0.201	*	0.318	0.318
43	0.135	0.359	0.036	0.976	0.007	*	0.059	*
44	0.901	0.057	0.095	0.022	0.037	*	0.825	0.036
45	0.430	0.008	0.165	*	0.100	*	0.133	0.675

P-values of Mann-Whitney-test, * $P < 0.001$.

Table IV. Fit and No-Fit for All Models and All Cosmonauts.

STATION	COSMONAUT	MODEL 31	MODEL 32	MODEL 33	MODEL 41	MODEL 42	MODEL 43	MODEL 44	MODEL 45
1	A	1	1	1	0	0	0	0	0
	B	0	0	0	0	0	0	0	0
	C	1	0	0	0	0	0	0	0
	D	1	1	1	1	0	1	1	1
	E	0	0	0	0	0	0	0	0
2	F	1	1	0	0	0	1	1	0
	G	0	0	0	0	0	0	0	0
	H	0	0	0	0	0	0	0	0
	I	1	1	1	0	0	1	1	1
	K	1	1	0	0	0	0	0	0
	L	1	1	0	1	0	1	0	0
	M	1	0	0	0	0	0	0	0
	N	1	0	0	1	1	1	1	1
	O	1	1	1	0	0	0	0	0
	P	1	0	0	1	1	1	1	1
	R	1	0	0	0	0	0	0	0
	S	1	1	0	1	1	1	1	1

but could not objectively guarantee the comparability of evaluations between different training sessions of a cosmonaut or even between different cosmonauts.

A factor-analytic verification seemed to be appropriate to compare the expert evaluations with a strictly mathematical one. Different factor-analytical approaches were tested and a common analysis over all available variables did not provide any reasonable results. The flight phase wise approaches promised to be more successful. Additionally, the final "docking contact" was analyzed separately. Within the system of expert coefficients of exactness, this moment was included in the "final approach" phase (Kt_5). For the three flight phases "stabilization2," "final approach," and "docking contact," factor models could be found reducing the large amount of raw parameters but still explaining most of the data variance.

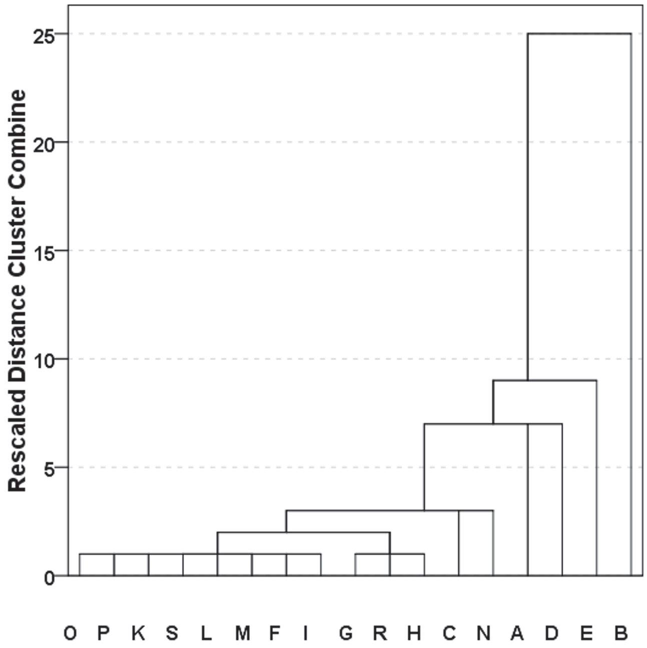

Fig. 6. Cluster dendrogram of factor-analytic performance values for each cosmonaut.

Dividing the Kt_5 into a "final approach" performance and a separate performance of the "docking contact" provided interesting results. The most striking seems to us that the performance during the final approach was not different between both space stations, but rather the separately evaluated docking contact moments were of significantly higher quality on the ISS (Fig. 4). As shown in Table II, correlation between original performance scores and factor-analytic scores for the "stabilization2" (Kt_4) and "final approach" (Kt_5) flight phases were either not statistically significant or of very low significance.

The expert performance evaluations of phase 5 (Kt_5), the "final approach" inclusive "contact," remain difficult to interpret. Good correlation was found between the common coefficient of the Russian standard expert evaluation (Kt) and the common factor-analytic coefficient of exactness (Kt_f). Also the reunified $Kt_{f\text{–}Kt5}$ correlated highly with Kt_f.

For a statistical verification of the found factor structures by means of confirmatory factor modeling, we reunified the "final approach" and the "docking contact" to be comparable to the

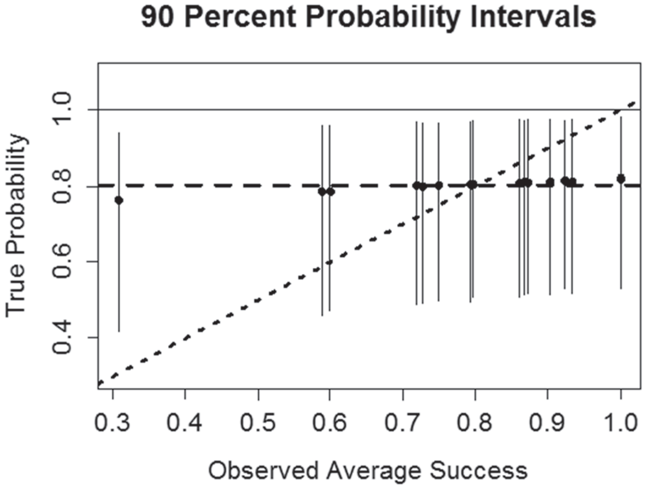

Fig. 7. Bayesian probability intervals of success for each cosmonaut. Dashed horizontal line: mean expected probability of successful docking for all cosmonauts; dotted diagonal line: individual probability of success.

Kt_5. A three-factor model was found for the "final approach" separately and a four-factor model could be confirmed for the "re-unified" flight phase 5. However, a separate factor, derived from the variables describing the contact moment, appears in the four-factor model. This suggests that the moment of contact is independent of the former final approach. The separate evaluation of the "docking contact" by the pass-fail criterion also supports our separate approaches with factor-analytic methods for the different flight phases, separating the contact moment. In our opinion, the Kt_5 coefficient especially needs more detailed analyses and, as concluded, improvements. However, in summary one can conclude that the expert evaluation by means of Kt's could be generally confirmed by factor-analytic verification.

In Table III, it is shown that these factor models are related to performance results assessed by the different indicators. For the group of data where a model fits, higher performance results were found. To us, this seems to be worth following up in future research.

The confirmatory FA model confirmation was different for individual cosmonauts. Table IV represents the individual pattern of fit and nonfit of the models for all cosmonauts. This could possibly be an approach to assess individual control styles in docking maneuvers. We assume that a model with fewer interrelations among the basic factors could describe a higher skill level of the operator. This should also be a topic for future research. For the use of these confirmatory FA models it will be necessary to ensure that the cosmonauts could run enough training flights so that the models are not rejected due to low numbers as happened with the first data.

Individual styles of docking control could also be assessed by means of cluster analysis using the factor-analytic flight phase wise performance evaluation. It could be shown that the control style was completely different between cosmonauts on the Mir station and on the ISS. Additionally, the styles among the Mir cosmonauts were nonhomogeneous. We interpret this again as an effect of an intensified docking training preflight using the onboard system which was used on ISS also for the PILOT experiment. This resulted in a more equalized performance as well as a more homogenized control style of the ISS cosmonauts.

The final aim of all docking training is to guarantee the docking success and, if possible, to predict the expected success quality and probability. For the evaluation of a docking training flight, in practice a strict data-based decision had to be made: 12 parameters had to be within defined safety ranges. Based on the pass-fail criterion, it is possible to calculate the conditional probability for success if the previous test flight was successful. By the extended Bayesian inference method of multilevel modeling, one can estimate the expected performance range. We have chosen a large probability range of 95% for a high likelihood of the predicted result; however, this results in larger deviation ranges (Fig. 7). There is a stringent conclusion that individuals with nearly 100% success during training flights are required to have an acceptable prediction for future docking success.

For future research it is desired that methods of performance evaluation are able to be repeatedly analyzed based on the whole training flight and on all available parameters, including all inputs from the control handles. Immediate onboard feedback is mandatory and was successful on Mir and ISS. However, for the use of new analytical methods established during the last few years, the provided data for a post analysis should be enlarged as the data transfer bandwidth from space is no longer a limiting factor.

Overall, the PILOT experiment demonstrated that the performance level of Russian cosmonauts in a mission relevant maneuver, the hand controlled docking of a spacecraft on a space station, was found to be significantly improved on the ISS in comparison to the Mir station. This can be interpreted as an enhancement of whole mission safety. In our opinion the main reasons are the increased number of docking training sessions (including the experimental sessions) and the increased number of flight tasks during a session. For future missions, a further increase in training tasks or even a special self-sufficient educational program could be useful for astronauts with less docking maneuver training prior to their flight. However, docking training over a period of, e.g., 3 yr does not appear to be necessary if the skill set is only needed at the end of a mission. Therefore, a training system that individually analyzes weaknesses and suggests adequate training sessions is desired.

ACKNOWLEDGMENTS

We would like to thank Karina Marshall-Bowman for her support in editing of the manuscript. We are thankful to the reviewers of this article, who helped us greatly to improve the text.

This work was continuously funded as a cooperative, joint project by ROSKOSMOS and DLR.

Authors and affiliations: Bernd Johannes, Ph.D., Institute of Aerospace Medicine, German Aerospace Center (DLR), Cologne, Germany; Vyacheslav Salnitski, Ph.D., and Alexander Dudukin, Ph.D., Institute for Biomedical Problems (IBMP) of the Russian Academy of Sciences, Russian Federation State Research Center, Moscow, Russia; and Lev Shevchenko and Sergey Bronnikov, Ph.D., S. P. Korolev Rocket and Space Corporation "Energia," Moscow, Russia.

REFERENCES

1. Albert J. Bayesian computation with R. New York: Springer; 2009. [Accessed 16 July 2015]. Available from https://www.infona.pl/resource/bwmeta1.element.springer-a2893eef-8822-3e6c-94f6-68ef2512aab5.
2. Albert J. Introduction to multilevel modeling. 2014 [Accessed February 2015]. Available from https://cran.r-project.org/web/packages/LearnBayes/index.html/MultilevelModeling.pdf.
3. Backhaus K, Erichson B, Plinke W, Weiber R. Multivariate analysemethoden, 8th ed. Berlin (NY): Springer; 1996.
4. Dudukin AV, Salnitski VP, Boritko YaC, Gushin VI, Vinohodova AG, et al. [Association between the person-unique individual behavior styles and the quality and reliability of operator's professional performance]. Aviakosm Ekolog Med. 2013; 47(3):10–19 [in Russian].
5. Johannes B, Salnitski V, Wittels P. Two psychophysiological scales for the description of psycho-physiological load. Proceedings of the 55th IAC Congress, 2004 Oct. 4–8; Vancouver, Canada. Reston(VA): AIAA; 2004.

6. Johannes B, Gaillard A. A methodology to compensate for individual differences in psychophysiological assessment. Biol Psychol. 2014; 96: 77–85.
7. Komotski RV, Salnitski VP. On the question of quality evaluation of ergonomic control processes. In: Tschernigovski VN, editor. Problemy kosmicheskoi biologii. Moscow (Russia): Nauka; 1977; 34:72–82 [in Russian].
8. Komotski RV, Minaev SA, Nechaev AP, Ryabov EV, Salnitski VP. Application of semi-natural modeling methods for the optimization of hand control systems. In: Tschernigovski VN, editor. Problemy kosmicheskoi biologii. Moscow (Russia): Nauka; 1977; 34:82–96 [in Russian].
9. Myasnickov VI, Bronnikov SV, Zhdanov OI. Psychological analysis and monitoring of crew performance. In: Dietlein FL, Pestov ID. Space biology and medicine. Vol. IV. Health, performance and safety of space crews. Joint U.S./Russian publication. Washington (DC): American Institute of Aeronautics and Astronautics; 2001:227–240.
10. Nechaev AP, Myasnikov VI, Stepanova SI, Isaev GF, Bronnikov SV. Some aspects of psychophysiological support of crewmember's performance reliability in space flights. Acta Astronaut. 2004; 54(10):749–754.
11. Nechaev AP, Myasnikov VI, Stepanova SI, Kozerenko OP, Isaev GF, Bronnikov SV. Methodological approach to study of cosmonauts errors and its instrumental support. Acta Astronautica. 1998; 42(1–8):331–338.
12. Nunnally JC, Bernstein IH. Psychometric theory, 3rd ed. New York: McGraw-Hill; 1994.
13. Salnitski VP. Evaluation and prognosis of reliability of cosmonauts professional work. [Otsenka I prognozirovanie nadezhnosti professionalnoi deyatelnosti kosmonavta] In: Myasnikov VI, Stepanova SI, Salnitski VP, Kozerenko OP, Nechaev AP. Problems of psychic asthenisation during long-term space flight. [Problema psihicheskoi astenisatsii v dlitelnom kosmicheskom polete.] Moscow (Russsia): Slovo; 2000:94–123 [in Russian].
14. Salnitski V, Bobrov A, Dudukin A, Johannes B. Reanalysis of operator's reliability in professional skills under simulated and real space flight conditions. Proceedings of the 55th IAC Congress. 2004 Oct 4–8; Vancouver, Canada. Reston (VA): AIAA; 2004.
15. Salnitski VP, Bobrov AF, Sheblanov VJu, Johannes B. Profession stress and its influence on operator's working reliability under long-term isolation. In: Grigoriev AI, Baranov VM, Buravkova LB, Voitkevich ND, Krugovyh VV, et al., editors. Life support problems in hermetic objects. Moscow: Slovo; 2001:162–163 [in Russian].
16. Salnitski 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.
17. Salnitski VP, Poljakov VV, Myasnikov VI, Shlykov JuV, Johannes B, Shevchenko LG. Psychodiagnostic complex trainer and its use in practice of manned space flight. XI Conference on Space Biology and Airspace Medicine; 1998 June 22–26. Moscow: Slovo; 1998; 2:185–186.