Characterizing Individual Differences in a Dynamic Stabilization Task Using Machine Learning

Vivekanand Pandey Vimal; Han Zheng; Pengyu Hong; Lila N. Fakharzadeh; James R. Lackner; Paul DiZio

INTRODUCTION: Being able to identify individual differences in skilled motor learning during disorienting conditions is important for spaceflight, military aviation, and rehabilitation.

METHODS: Blindfolded subjects ($N = 34$) were strapped into a device that behaved like an inverted pendulum in the horizontal roll plane and were instructed to use a joystick to stabilize themselves across two experimental sessions on consecutive days. Subjects could not use gravitational cues to determine their angular position and many soon became spatially disoriented.

RESULTS: Most demonstrated minimal learning, poor performance, and a characteristic pattern of positional drifting during horizontal roll plane balancing. To understand the wide range of individual differences observed, we used a Bayesian Gaussian Mixture method to cluster subjects into three statistically distinct groups that represent Proficient, Somewhat Proficient, and Not Proficient performance. We found that subjects in the Not Proficient group exhibited a suboptimal strategy of using very stereotyped large magnitude joystick deflections. We also used a Gaussian Naive Bayes method to create predictive classifiers. As early as the second block of experimentation (out of ten), we could predict a subject's final group with 80% accuracy.

DISCUSSION: Our findings indicate that machine learning can help predict individual performance and learning in a disorienting dynamic stabilization task and identify suboptimal strategies in Not Proficient subjects, which could lead to personalized and more effective training programs.

KEYWORDS: machine learning, dynamic balance, vehicle control, spatial disorientation, motor skill learning, vestibular system, somatosensation, spaceflight analog.

Vimal VP, Zheng H, Hong P, Fakharzadeh LN, Lackner JR, DiZio P. *Characterizing individual differences in a dynamic stabilization task using machine learning.* Aerosp Med Hum Perform. 2020; 91(6):479–488.

Detecting individual differences and using them to make predictions of performance and learning are important for development of motor skills and rehabilitation training. In previous work, we studied learning of a dynamic balancing task in a spaceflight analog condition.[12,19,21] Blindfolded subjects controlled a Multi-Axis Rotation System device (MARS) that was programed to behave like an inverted pendulum.[13,20] Subjects used a joystick to stabilize themselves around the balance point. When they balanced in a Vertical Roll Plane (**Fig. 1**) and tilted relative to the gravitational vertical they received information about their angular position relative to gravity from their otolith organs and somatosensory receptors, and information about their angular velocity from their semicircular canals and somatosensory system. In this circumstance, subjects showed robust learning across many performance metrics. In contrast, in a spaceflight analog condition, subjects balanced in the Horizontal Roll Plane where they no longer tilted relative to the gravitational vertical and therefore had no gravitational cues about their angular position and had to rely on motion cues. Collectively, these subjects showed poor performance, minimal learning, and a characteristic pattern of positional drifting.[20] Similar results were found when subjects tried to dynamically stabilize about a Vertical Yaw Axis, where they also could not rely on gravitational cues.[21] However, within

From Brandeis University, Waltham, MA.

This manuscript was received for review in November 2019. It was accepted for publication in March 2020.

Address correspondence to: Vivekanand Pandey Vimal, Ashton Graybiel Spatial Orientation Laboratory, MS 033, Brandeis University, Waltham, MA 02245-9110; somde@brandeis.edu.

DOI: https://doi.org/10.3357/AMHP.5552.2020

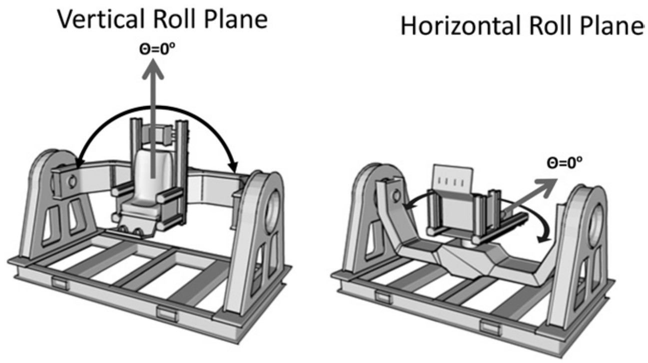

Fig. 1. The multiaxis rotation device (MARS) was programmed with inverted pendulum dynamics in the vertical roll axis (left) and the horizontal roll axis (right). Straight grey arrows represent the direction of balance.

these spaceflight analog conditions, when we examined the performance of individual subjects, we were surprised to find that some showed major improvements while others became worse across multiple performance measures.[18] This pattern led us to explore three core questions that guide this paper. First, can we group subjects into clusters with statistically distinct proficiencies? Second, does each cluster have its own signature pattern and can this give us insights into why certain subjects learn and some continue to perform poorly? Finally, how early can we predict the final performance of a given subject?

METHODS

Subjects

There were 34 healthy adult subjects (18 women and 16 men, 20.4 ± 2.0 yr old) who had no prior experience in the Multi-Axis Rotation System (MARS) and gave written consent to participate in the experiment as approved by the Brandeis Institutional Review Board. Of the 34 subjects, 14 were recruited specifically for this study, 10 were from the Control group (Horizontal Roll condition) of the Vimal et al. 2017 study[19] and 10 were from the Control group (Horizontal Roll condition) of the Vimal et al. 2019 study[18] to ensure a large enough group to allow meaningful performance clustering.

Equipment

The MARS was programmed with inverted pendulum dynamics about a horizontal roll axis as shown in Fig. 1. MARS dynamics were governed by the equation, $\ddot{\theta} = k_P sin\theta$, where θ is the angular deviation from the direction of balance, DOB (in degrees), and k_P is the pendulum constant. As in our previous work, we used a pendulum constant of 600 deg · s^{-2} (≈0.52 Hz).[19–21] We programmed 'crash' limits that restricted the angular range of the MARS to ± 60 deg from the direction of balance. Angular velocity was limited to ± 300 deg · s^{-1}, and angular acceleration to ± 180 deg · s^{-2}. Further details are available in Panic et al.[13] At every time step (~0.02 s), a velocity increment proportional to the joystick deflection was added to the MARS velocity and computed by a Runge-Kutta RK4 solver[9] to calculate the new MARS angular position and velocity. The latency between a joystick deflection and a change in MARS angular velocity was 30 ms over the observed range of MARS spectral power of 0 to ≈0.75 Hz.

Procedure

Subjects were secured in the MARS with a five-point harness, a lap belt, lateral support plates and foot straps (Fig. 1). The MARS was configured for Horizontal Roll for all subjects. Their heads were stabilized using a U-shaped frame cushioned with foam that was attached to the MARS. To prevent visual or auditory cues, they were blindfolded and wore earplugs and noise cancelling headphones that played white noise. A Logitech Freedom 2.4 cordless joystick was attached to the right arm rest and a "kill switch" was attached to the left arm rest that the subject could press to stop the experiment. No subject ever used the kill switch.

The experimental design required each subject to complete 40 balancing trials divided equally over 2 consecutive days. Prior to data collection subjects watched a video of a person balancing the MARS in the Horizontal Roll Plane and of the MARS reaching the "crash boundaries" at ± 60 deg from the DOB (direction of balance) and then resetting. They were told that the MARS behaved like an inverted pendulum and to use the joystick to balance it at the DOB while also minimizing oscillations. Subjects were also familiarized with the symptoms of motion sickness using the Graybiel Diagnostic Criteria.[8] After signing consent forms, they were secured in the MARS, given a blindfold, earplugs, and noise cancelling headphones that played white noise. The 10 subjects from the Vimal et al. 2019 study[18] had also been given conscious strategies that the other 24 subjects did not receive. We found no statistical differences between those subjects and the 10 from Vimal et al. 2018[21] study or the 14 new subjects tested here.

Subjects heard an auditory "begin" at the onset of a trial. Whenever they reached the crash boundaries, they heard "lost control, resetting", and during the reset the joystick was disabled as the MARS automatically reset to the start position at a rate of 5 deg · s^{-1}. Once at the reset position, which was always 0 deg, they heard an auditory "begin" command and the joystick was simultaneously enabled. Subjects participated in two sessions conducted on consecutive days. On each day they underwent 5 blocks of 4 trials, with each trial consisting of 100 cumulative seconds of balancing, excluding the reset times after crashes, or a total elapsed time of 150 s. After every 4 trials subjects were brought to an upright orientation and were given a 2-min break during which they were questioned about any symptoms of motion sickness. They were given no verbal feedback about their performance.

Statistical Analysis

Data from the reset phase following a crash when subjects had no control over the MARS were not included in the analysis. We applied a zero-phase, 5-pole high pass Butterworth filter with a cutoff frequency of 5 Hz on the MARS angular position and velocity data, and joystick deflections, all of which were sampled at 20.7 ± 1.1 ms (approximately 50 Hz). Unless otherwise noted, each of the following measures described below was calculated in every trial and then averaged across the four trials in a block.

MARS Performance. MARS performance was quantified by calculating the average MARS angular position ($Mean_{MARS}$) and the standard deviation of MARS angular position (STD_{MARS}). The frequency of crashes (Crashes) was found by counting the number of crashes in a trial and then dividing by the trial duration. The average MARS angular deviation from the direction of balance ($|Mag|_{Pos}$), and the average magnitudes of MARS velocity ($|Mag|_{Vel}$) and acceleration ($|Mag|_{Accel}$) were calculated by taking the mean of the absolute value of each measure. We quantified the rate of drift (DriftRate) by tracking the center of loops in the MARS position versus velocity phase plots (see **Fig. 2**; for color see online at https://doi.org/10.3357/AMHP.5552.2020). The loops were identified in the periods without crashes (balance time between consecutive crashes) in each trial. Loop "quadrants" were defined as points in each loop with unique velocity and acceleration attributes that could be visualized by drawing two orthogonal lines intersecting at the loop center. The horizontal positions of the points in the loops with maximum velocity were used to identify the "centers." The rate of drift (DriftRate) was calculated by fitting a regression line to the hemiloop center positions (which was the mean absolute value of the loop center velocity) versus time and taking the mean absolute values of the slopes for individual segments.

Joystick Commands. Joystick commands were quantified by calculating the average of the absolute value of joystick deflections ($|Mag|_{Joy}$), which could vary from +1 to -1 for full deflection. Intermittency of joystick deflections ($\%Zero_{Joy}$) was determined by finding the percentage of data points where the joystick deflection was $\leq \pm 1\%$ of its maximum amplitude. Anticipatory joystick deflections (%Anticipatory; Fig. 2) were defined as those that removed energy from the MARS by decelerating it as it was moving toward the DOB. We calculated the percentage of anticipatory joystick deflections by finding the number of data points where the MARS angular position and joystick deflection had opposite signs to the MARS angular velocity and then dividing by the total number of data points. While anticipatory joystick deflections can help stabilize the MARS, they are often used when poor control has led to large velocities near the balance point. When subjects learn to stabilize the MARS, the percentage of anticipatory joystick deflections decreases. Destabilizing joystick deflections were defined as those that add energy to the MARS, accelerating it away from the DOB. We calculated the percentage of destabilizing joystick deflections (%Destab; Fig. 2) by finding the number of data points where the MARS angular position and velocity, and the joystick deflection all had the same sign and then dividing by the total number of data points.

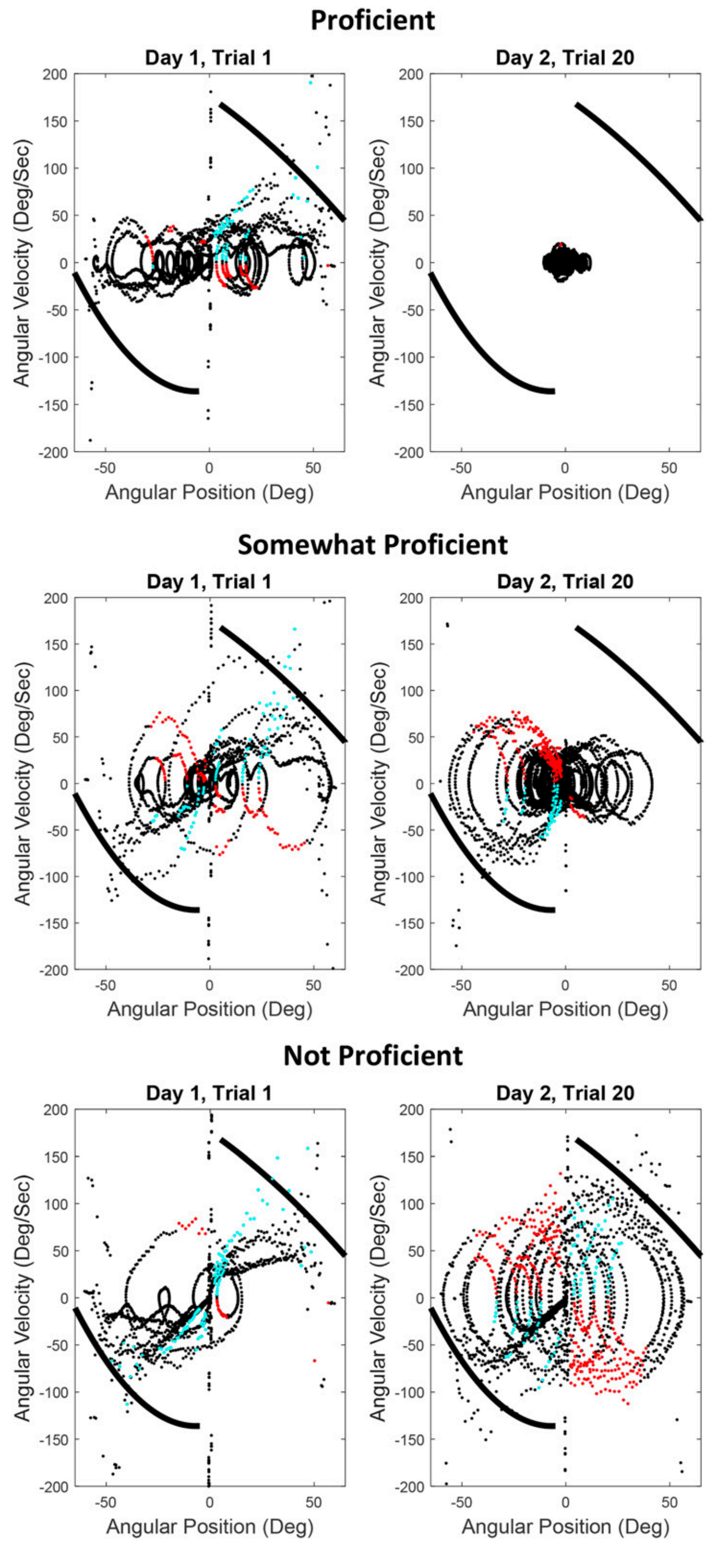

Fig. 2. Position-velocity phase plots from representative subjects from the Proficient (top), Somewhat Proficient (middle) and Not Proficient (bottom) groups. Graphs on the left are from the first trial on Day 1 and graphs on the right are from the final trial on Day 2. The thick lines represent empirically determined crash boundaries from which joystick deflections cannot lead to recovery. Light grey points (light blue points online) are destabilizing joystick deflections and dark grey points (red points online) are anticipatory joystick deflections. (See color online figure at https://doi.org/10.3357/AMHP.5552.2020.)

The average instantaneous phase difference between joystick deflections and MARS angular position is an index of the regularity of joystick commands in relation to MARS motion. To calculate it we used an approach introduced by Park et al.[14] First, the data from joystick deflections and MARS angular position were detrended to remove the drift by subtracting the moving mean average, with a sliding window of 1 s, from the raw data. Next, the angular component of the Hilbert transform for the detrended MARS angular position and joystick deflection were calculated and were subtracted to obtain the instantaneous phase difference. To obtain the standard deviation of this phase difference, we used the circular standard deviation function. For every trial, we calculated the instantaneous phase for the segments of data between crashes and then averaged them. Sometimes after a crash, a subject initially would not deflect the joystick so we only analyzed the data after the joystick became deflected more than 1% of its maximum. A smaller average phase difference signifies less variability in joystick deflections in relation to MARS angular position.

Stabilogram Diffusion Function. The Stabilogram Diffusion Function was developed by Collins and DeLuca to model human quiet stance balancing.[3,4] It involves calculating the mean-squared displacement (MSD) of sway as a function of different time intervals:[4]

$$MSD = <\Delta\theta^2>_{\Delta t} = \frac{\sum_{i=1}^{N-m}(\Delta\theta_i)^2}{(N-m)}$$

where θ is MARS angular position, Δt is the time interval, N is the total number of data points, and m is the number of data points for the total time interval. Like Collins and DeLuca,[4] we set the maximum time interval to 10 s and calculated the MSD over time intervals that did not have any crashes. The Stabilogram Diffusion Function allowed Collins and DeLuca to characterize postural quiet stance balancing over two regimes: the short-term random walk regime and the longer term corrective adjustment regime. In our prior work, we applied the Stabilogram Diffusion Function to dynamic stabilization in the MARS.[19–21] We found that subjects balancing in the Vertical Roll Plane rapidly improved their balance control; however, subjects in the Horizontal Roll Plane showed delayed learning and persistent drifting in the long term.

To find the short- and long-term regimes, we first calculated the Hurst scaling exponent (H) using the equation:

$$<\Delta\theta^2> = \Delta t^{2H}$$

H was then obtained by plotting the log of MSD vs. the log of Δt, calculating its slope, and then dividing by 2. Described in our previous papers in greater detail,[19–21] H is important because when it changes from $> 0.5\ \mathrm{deg}^2 \cdot \mathrm{s}^{-1}$ to < 0.5, it represents a transition between the short- and long-term regimes that is called the "critical point" (CP). When H is > 0.5 it means the MARS is continuing to move in the same direction, when it is $= 0.5$ the MARS motion is Brownian, and when it is < 0.5 it is antipersistent. We also calculated the diffusion coefficient (D) of the SDF, which represents the rate of change of the MSD:

$$<\Delta\theta^2> = 2D\Delta t$$

The short-term diffusion coefficient (D_S) is found by taking one-half of the slope between $\Delta t = 0$ s and the critical point on the standard MSD plot. The long-term diffusion coefficient (D_L) corresponds to one-half of the slope between the critical point and $\Delta t = 10$ s. If $D_L > 0\ \mathrm{deg}^2 \cdot \mathrm{s}^{-1}$, the MSD continues to increase in long time intervals. To quantify the overall energy of the long-term regime ($Mean_{MSD}$), we calculated the average MSD between the CP and $\Delta t = 10$ s. Both the $Mean_{MSD}$ and DriftRate are sensitive to the average rate of positional drifting, whereas D_L represents the long-term rate of change in the mean-squared displacement.

Machine Learning

Our first task was to identify clusters of subjects with distinct performance proficiencies at the end of Day 2 after 40 trials. We grouped individual subjects into one of three clusters using their data from the final block of Day 2, and then verified that the three clusters represented 'Proficient', 'Somewhat Proficient', and 'Not Proficient' performance (**Fig. 3** and **Table I**). Next, we applied the Gaussian Naïve Bayes method to build models capable of predicting the final performance of an individual using his/her performance data in early blocks. In this step, we reduced the redundancy of the features (i.e., performance measures) and did feature engineering to create new features representing the learning of subjects.

Discovering Proficiency Clusters. In clustering analysis, we first selected a list of metrics that are essential to measure learning and performance based on our prior work.[18,19,21] We found that subjects in the Vertical Roll condition learned to improve performance by reducing the standard deviation of angular position (STD_{MARS}), the number of crashes (Crashes), the

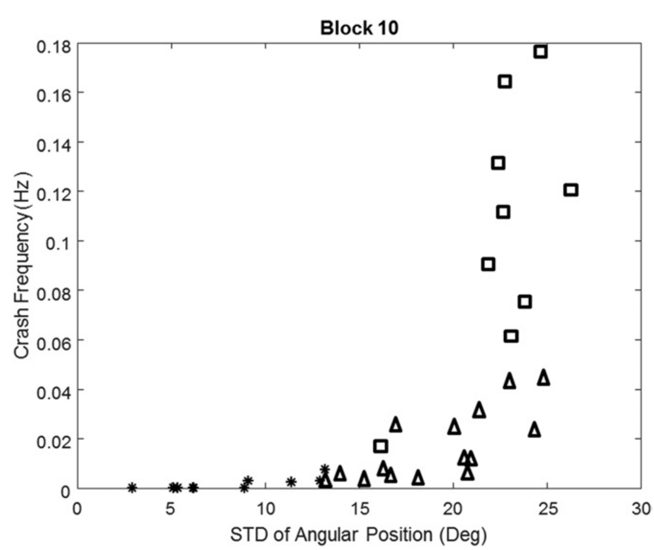

Fig. 3. Data from all 34 subjects are plotted in a graph of Crash Frequency vs. Standard Deviation of Angular Position. These measures were chosen because they are good indicators of performance. The stars (*) represent Proficient subjects, the triangles (▲) represent Somewhat-Proficient and the squares (□) represent Not-Proficient subjects.

Table I. Paired *t*-Tests Within Groups.

METRIC	PROFICIENT (N = 10) DAY1BLOCK1	PROFICIENT (N = 10) DAY2BLOCK5	SOMEWHAT PROFICIENT (N = 15) DAY1BLOCK1	SOMEWHAT PROFICIENT (N = 15) DAY2BLOCK5	NOT PROFICIENT (N = 9) DAY1BLOCK1	NOT PROFICIENT (N = 9) DAY2BLOCK5
MARS PERFORMANCE						
STD_{MARS} (deg)	21.3	**8.1*** **	22.0	**19.2***	21.1	22.9
Crashes (Hz)	0.09	**0.002*** **	0.13	**0.02*** **	0.23	**0.11** **
$\|Mag\|_{Pos}$ (deg)	16.7	**6.5*** **	16.4	16.2	14.8	*17.8**
$\|Mag\|_{Vel}$ (deg/s)	18.3	**6.9*** **	24.5	**17.6*** **	30.4	36.4
$\|Mag\|_{Accel}$ (deg/s^2)	71	**29*** **	86	**66** **	100	*163***
DriftRate (deg/s)	4	**0.25*** **	5	**1*** **	5.5	4.5
JOYSTICK COMMANDS						
$\|Mag\|_{Joy}$	0.22	**0.08*** **	0.27	0.23	0.28	*0.50***
$\%Zero_{Joy}$	33	**53** **	31	30	35	*15***
%Destab	1.6	**5e-4*** **	3.2	**0.6*** **	6.8	4.8
%Anticipatory	3.4	**0.2*** **	5.1	**2.8***	5.2	*14***
STD_{Joy_Pos}	57.4	**67.8** **	53.7	55.8	52.2	40.2
STABILOGRAM-DIFFUSION FUNCTION						
D_S (deg^2/s)	111	**21*** **	210	**119*** **	69	354
D_L (deg^2/s)	29.5	**5.0*** **	28.7	24.2	45.4	32.7
$Mean_{MSD}$ (deg^2)	493	**112*** **	607	**453***	646	797

* $P < 0.05$; ** $P < 0.01$; and *** $P < 0.001$. Bolded values represent improving performance whereas italicized values represent worsening performance.

overall short term energy of the system (D_S), and the magnitude of angular velocity and acceleration ($|Mag|_{Vel}$, $|Mag|_{Accel}$). They did so by reducing the magnitude of joystick deflections ($|Mag|_{Joy}$) and the number of destabilizing and anticipatory joystick deflections (%Destab, %Anticipatory) while increasing the percentage of intermittent joystick deflections (%Zero). In the Horizontal Roll condition, subjects displayed a pattern of positional drifting not seen in the Vertical Roll condition, so we included DriftRate and D_L with the idea that good performers may learn to decrease them. Finally, many poor performers used very stereotyped joystick deflections which we identified as being suboptimal (STD_{Joy_Pos}). Next, we designed a Bayesian Gaussian Mixture model (BGMM)[2] to group subjects into three clusters because our analysis showed that there were three statistically distinct clusters. Our preliminary examination revealed that the three clusters corresponded to: subjects who improved across many metrics, subjects who became worse across many metrics, and subjects who had a mixture of these two. To evaluate how different two clusters were from each other, we developed a new score, termed MSilhouette, by incorporating the Mahalanobis distance[10] into the Silhouette score.[16] Unlike the traditional Silhouette score, our MSilhouette score uses the Mahalanobis distance to handle the correlations between metrics, and then effectively reduces the influence of redundancies among the chosen metrics. We used the MSilhouette score in the random permutation test, MRPT, (see online **Appendix** sections 1.1,2,3 for details; https://doi.org/10.3357.AMHP.5552sd.2020) which showed the obtained subject clusters are pair-wise significantly different. The MRPT produced three P-values, one for each of the comparisons between the Proficient, Somewhat Proficient, and Not-Proficient groups. Because we conducted three comparisons, we used a Bonferroni adjusted value of 0.0167. All of the P-values were less than 0.0167, so we concluded that the clusters were all statistically distinct.

To justify the labels of the obtained clusters, we visually examined the clusters to see whether they truly reflected training performances. Fig. 3 shows that the clusters faithfully represent three different levels of training outcomes: Proficient (smaller STD_{MARS} and Crashes), Somewhat Proficient, or Not Proficient (larger STD_{MARS} and Crashes). **Table II** summarizes the independent t-test results confirming that the majority of the chosen metrics are statistically distinct between clusters.

Feature Reduction and Engineering. Before building the predictive classifier, we reduced the number of features (i.e., metrics) from the original 12 (STD_{MARS}, Crashes, D_S, $|Mag|_{Vel}$, $|Mag|_{Accel}$, $|Mag|_{Joy}$, %Destab, %Anticipatory, %Zero, DriftRate, D_L, STD_{Joy_Pos}) to avoid redundancy that could lead to potential overfitting. We did this by first taking out four features and then using the remaining 8 to perform the clustering analysis described above. When the clusters were still statistically distinct, we assumed that the 4 removed features were redundant. We rotated through all combinations and kept all of the sets of 8 features that met the criteria of having a P-value less than 0.0015. We then merged those lists together to obtain the final list: STD_{MARS}, Crashes, D_S, $|Mag|_{Vel}$, %Destab, %Anticipatory, DriftRate, D_L.

We trained the predictive classifier using the selected metrics listed above and further explored the learning demonstrated by subjects from block to block. The learning is reflected in changes of the metrics over blocks, which indicate whether a given subject is able to improve from his or her previous experience. Therefore, we designed the learning features for each block as the changes of the above chosen metrics, i.e., the difference of a chosen metric in the current block and its means in previous blocks (refer to sections 4 and 5 of the online **Appendix I**; https://doi.org/10.3357/AMHP.5552sd.2020). The number of the engineered features at the n-th block is $M(n-1)n/2$, which will quickly exceed the number of samples as n grows and causes over-fitting in training our classifier. To tackle this

Table II. Independent *t*-Tests Between Groups (Day 2 Block 5).

METRIC	PROFICIENT	NOT PROFICIENT	PROFICIENT	SOMEWHAT PROFICIENT	SOMEWHAT PROFICIENT	NOT PROFICIENT
MARS PERFORMANCE						
STD_{MARS} (deg)	8.1	22.9**	8.1	19.2**	19.2	22.9
Crashes (Hz)	0.002	0.11**	0.002	0.02*	0.02	0.11**
$\lvert Mag\rvert_{Pos}$ (deg)	6.5	17.8**	6.5	16.2**	16.2	17.8
$\lvert Mag\rvert_{Vel}$ (deg/s)	6.9	36.4**	6.9	17.6**	17.6	36.4**
$\lvert Mag\rvert_{Accel}$ (deg/s^2)	29	163**	29	66**	66	163**
DriftRate (deg/s)	0.25	4.5**	0.25	1*	1	4.5**
JOYSTICK COMMANDS						
$\lvert Mag\rvert_{Joy}$	0.08	0.50**	0.08	0.23**	0.23	0.50**
%$Zero_{Joy}$	53	15**	53	30**	30	15*
%Destab	5e-4	4.8**	5e-4	0.6*	0.6	4.8**
%Anticipatory	0.2	14**	0.2	2.8*	2.8	14**
STD_{Joy_Pos}	67.8	40.2**	67.8	55.8**	55.8	40.2**
STABILOGRAM-DIFFUSION FUNCTION						
D_S (deg^2/s)	21	354**	21	119*	119	354**
D_L (deg^2/s)	5.0	32.7**	5.0	24.2**	24.2	32.7
$Mean_{MSD}$ (deg^2)	112	797**	112	453**	453	797*

* $P < 0.05$; ** $P < 0.01$; and *** $P < 0.001$. Reported P levels have been Bonferroni corrected.

problem, we applied two feature selection methods: a tree-based feature selection method[5] and the L1 regularization,[6] and obtained better performance in the leave-one-out cross validation test.

Training the Predictive Classifier. One of our main goals is to build a robust classifier capable of predicting the final performance (proficient, somewhat proficient, not proficient) of individual subjects as early as possible in their 2-d experiments. Due to the difficulty and time required to run 2-d experiments, we only had 34 participants. For such a small number of participants, we used the leave-one-out cross validation method[17] to ensure that we trained a robust classifier for accurately predicting training results. Each time 33 participants were used to build a classifier, which was then used to predict the training performance of the left-out participant. This procedure was repeated 34 times and the accuracies were averaged and summarized in **Fig. 4**. We found that the Gaussian Naive Bayes method with a uniform class prior outperformed other conventional methods, such as logistic and Lasso regression, in the leave-one-out cross validation.

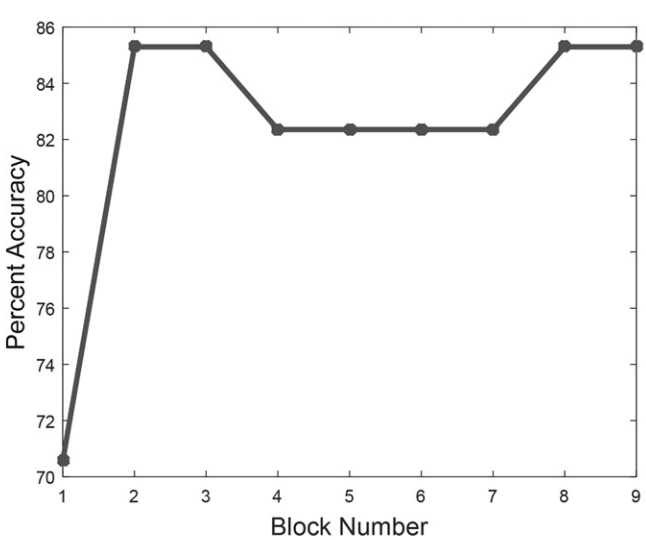

Fig. 4. The accuracy of predictions to determine final performance in Block 10 on day 2.

RESULTS

Following Collins and De Luca,[3] we first averaged the SDFs across the four trials in each block for each subject, and then calculated the short-term and long-term diffusion coefficients, and the final MSD. We then averaged the values across all participants. For all other measures, we first calculated them for each trial and then averaged them across trials in a block, and finally averaged them across all subjects.

To test for learning within groups, we carried out paired two-tailed *t*-tests between Block 1 of Day 1 to Block 5 of Day 2. To test for differences between groups, we performed independent two-tailed *t*-tests on Block 5 of Day 2 and applied the Bonferroni correction where, because each group was compared twice, we adjusted the P value to be less than 0.025. In the tables, statistical significance is denoted with symbols, where '*' was assigned when $P < 0.05$ and '**' when $P < 0.01$, and '***' when $P < 0.001$. Bonferroni corrections are a robust way to prevent Type I errors due to multiple comparisons per variable but are less efficient for correcting inflated chances of reporting coincidental results as real in comparisons of multiple variables per comparison. Our approach to maintaining the Type I error rate of multivariate hypothesis tests was to report all comparisons made, rather than picking only the significant ones. Under this condition, there is an extremely small likelihood that the entire pattern of results across all variables reported in Table II is coincidental.

The 34 subjects were clustered into three statistically distinct categories based on their data from the final block (Block 5) of the second day. To determine whether these clusters reflected differences in performance we plotted the data for two key metrics: the standard deviation of angular position and the crash frequency, which confirmed that the clusters define Proficient, Somewhat Proficient, and Not Proficient performers. Table I shows significant group differences across multiple metrics using independent *t*-tests.

To determine how early we could reliably predict final performance, we used the Gaussian Bayes Method. We found that by Block 2 of Day 1 we could reliably predict with 80% accuracy a subject's final group.

The Proficient Group (*N* = 10)

Table I shows that these subjects had significant learning across all metrics when comparing Block 1 of Day 1 to Block 5 of Day 2. In the MARS Performance category, they were able to decrease the standard deviation of angular position (STD_{MARS}), the frequency of crashes (Crashes), the magnitudes of MARS angular position ($|Mag|_{Pos}$), velocity ($|Mag|_{Vel}$), and acceleration ($|Mag|_{Accel}$), and the rate of positional drifting (DriftRate). In the Joystick Command category, they learned to decrease the magnitude of joystick deflections ($|Mag|_{Joy}$), the percentage of destabilizing joystick deflections (%Destab), the percentage of anticipatory joystick deflections (%Anticipatory), and to increase the intermittency of joystick deflections (%$Zero_{Joy}$) and the variability in the phase difference between joystick deflection and angular position (STD_{Joy_Pos}). In the Stabilogram-Diffusion Function category, the proficient subjects significantly decreased the short-term diffusion coefficient (D_S), the long-term diffusion coefficient (D_L) and the long-term mean-squared displacement ($Mean_{MSD}$).

The Somewhat-Proficient Group (*N* = 15)

For MARS Performance, these subjects significantly decreased the standard deviation of angular position (STD_{MARS}), the frequency of crashes (Crashes), the magnitudes of MARS angular velocity ($|Mag|_{Vel}$) and acceleration ($|Mag|_{Accel}$), and the rate of positional drift (DriftRate). For Joystick Commands, they decreased the percentage of destabilizing (%Destab) and anticipatory joystick deflections (%Anticipatory). In the Stabilogram-Diffusion Function category, they decreased the short term diffusion coefficient (D_S) and the average final mean-squared displacement ($Mean_{MSD}$).

The Not-Proficient Group (*N* = 9)

Subjects in the Not-Proficient Group improved on only one metric: they learned to decrease the frequency of crashes (Crashes). All other statistically significant changes indicated deteriorating performance. In the MARS Performance category, they statistically increased the magnitudes of both MARS angular position ($|Mag|_{Pos}$) and acceleration ($|Mag|_{Accel}$). In the Joystick Command category, they increased the magnitude of joystick deflections ($|Mag|_{Joy}$), and the percentage of anticipatory joystick deflections (%Anticipatory) and decreased the percentage of intermittent joystick deflections (%$Zero_{Joy}$).

The Proficient vs. Not-Proficient Group

Comparing Day 2 Block 5 data for the Proficient and Not-Proficient groups using independent *t*-tests showed highly significant differences (Table II). In the MARS Performance category, the Not-Proficient group had a 2.8 times greater standard deviation of angular position (STD_{MARS}), 55 times greater frequency of crashes (Crashes), 2.7 times greater magnitude of MARS angular position ($|Mag|_{Pos}$), 5.3 times greater velocity ($|Mag|_{Vel}$), 5.6 times greater acceleration ($|Mag|_{Accel}$), and 18 times greater rate of positional drifting (DriftRate). In the Joystick Command category, the Not-Proficient group had a 6.3 times greater magnitude of joystick deflections ($|Mag|_{Joy}$), 3.5 times lower intermittency of joystick deflections (%$Zero_{Joy}$), 100 times greater percentage of destabilizing joystick deflections (%Destab), 70 times greater percentage of anticipatory joystick deflections (%Anticipatory), and 1.7 times less variability in the phase difference of joystick deflection and MARS angular position (STD_{Joy_Pos}). In the Stabilogram-Diffusion Function category, the Not-Proficient group had a 16.8 times greater short-term diffusion coefficient (D_S), 6.5 times greater long-term diffusion coefficient (D_L), and 7 times greater mean-squared displacement ($Mean_{MSD}$).

The Proficient vs. Somewhat-Proficient Group

We found highly significant differences between the Proficient and Somewhat-Proficient groups across all metrics using independent *t*-tests comparing Block 5 data of Day 2 (Table II). In the MARS Performance category, the Somewhat-Proficient group had a 2.4 times greater standard deviation of angular position (STD_{MARS}), 10 times greater frequency of crashes (Crashes), 2.5 times greater magnitude of MARS angular position ($|Mag|_{Pos}$), 2.6 times greater velocity ($|Mag|_{Vel}$), 2.3 times greater acceleration ($|Mag|_{Accel}$) and 4 times greater rate of positional drift (DriftRate). In the Joystick Command category, the Somewhat-Proficient group had 3 times greater magnitude of joystick deflections ($|Mag|_{Joy}$), 1.8 times less the intermittency of joystick deflections (%$Zero_{Joy}$), more than 100 times greater percentage of destabilizing joystick deflections (%Destab), 14 times greater percentage of anticipatory joystick deflections (%Anticipatory), and 1.2 times less variability in the phase difference between joystick deflection and MARS angular position (STD_{Pos_Joy}). In the Stabilogram-Diffusion Function category, the Somewhat-Proficient group had a 5.6 times greater short-term diffusion coefficient (D_S), 4.8 times greater long-term diffusion coefficient (D_L), and 4 times greater long-term mean-squared displacement ($Mean_{MSD}$).

The Somewhat-Proficient vs. Not-Proficient Group

Significant differences between the Somewhat-Proficient and Not-Proficient groups were present for their Day 2 Block 5 data (Table II). In the MARS Performance category, the Not-Proficient group had 5.5 times greater frequency of crashes (Crashes), 2 times greater velocity ($|Mag|_{Vel}$), 2.5 times greater acceleration ($|Mag|_{Accel}$), and 4 times greater drift rate of angular position (DriftRate). In the Joystick Command category, the Not-Proficient group had 2 times greater magnitude of joystick deflections ($|Mag|_{Joy}$), 2 times less intermittency of joystick deflections (%$Zero_{Joy}$), 8 times greater percentage of destabilizing joystick deflections (%Destab), 5 times greater percentage of anticipatory joystick deflections (%Anticipatory), and 1.4 times greater variability in the phase difference between joystick deflection and MARS angular position (STD_{Joy_Pos}). In the Stabilogram-Diffusion Function category, the Not-Proficient group had a 3 times greater short-term diffusion coefficient (D_S) and 1.8 times greater long-term mean-squared displacement ($Mean_{MSD}$).

DISCUSSION

Being able to identify, classify, and predict individual differences in skilled motor learning is important for optimizing training as well as rehabilitation practices. Recent advances in machine learning techniques are helping accomplish these goals; however, they primarily rely on humans first labeling 'ground truths' (e.g., Proficient, Somewhat-Proficient, and Not-Proficient) to every subject in the training data set. This approach is difficult for complex real-life tasks where subjects can use many different strategies and can optimize different performance metrics. In our task, blindfolded subjects balance in the MARS in the Horizontal Roll plane where they do not have access to relevant positional gravitational cues from the otoliths and somatosensory receptors and can only use motion cues detected by the semicircular canals and somatosensory receptors. Because this condition is very disorienting, subjects develop a variety of different strategies. For example, one subject may be able to stabilize with relatively small oscillations of low velocity about the DOB but occasionally may make a destabilizing joystick deflection that causes a crash. Another subject may be able to avoid the crash boundaries but have large positional oscillations at high velocities. Which of these subjects is better? Do they belong to the same group? These are questions that cannot be easily answered and are why we did not approach our balancing paradigm with a machine learning approach of using data that had preassigned labels. Instead, we clustered subjects into three statistically distinct groups based on their final block of experimental trials on the second day.

To verify whether these clusters had meaning, we plotted all 34 subjects on a graph using two important indicators of performance, positional variability (STD_{MARS}) and the frequency of hitting the crash boundaries (Crashes). Fig. 3 shows that one cluster had infrequent crashes and small standard deviations of angular position (the Proficient group), another had much larger values (the Not-Proficient group), and another was between these two (the Somewhat-Proficient group). We confirmed the validity of these clusters by performing independent *t*-tests between groups, which showed significant differences across multiple metrics (Table I). After reducing redundant features and performing feature engineering, we used the Gaussian Naive Bayes method to create predictive classifiers. Fig. 4 shows that following Block 2 of Day 1 we can predict a given subject's final cluster with 80% or greater accuracy. One application for such an early detection of final performance is that those subjects who are predicted to perform poorly can receive specialized training such as that described in Vimal et al.,[18] which enhances performance in a spaceflight analog condition.

Clustering gives insights into the performance and learning of each group. The Proficient group showed statistically significant learning across all measures (Table I). They did this by learning to use smaller joystick deflections ($|Mag|_{Joy}$) and decreasing the percentages of anticipatory (%Anticipatory) and destabilizing (%Destab) joystick deflections while increasing the intermittency of the joystick deflections (%$Zero_{Joy}$). Reducing anticipatory joystick deflections reflects proficient learning because while anticipatory joystick deflections can stabilize the MARS, they are often used when poor control has led to large velocities near the balance point. By contrast, the Not-Proficient group only learned to decrease the frequency of crashes, otherwise all other statistically significant changes reflected worsening performance. How is it possible that these subjects could reduce the frequency of crashes when they did not improve in any other metric? **Fig. 5** shows representative trial data from a Not-Proficient subject (top) and a Proficient subject (bottom). It reveals that the Not-Proficient subjects tended to use a non-optimal strategy of very large magnitude joystick deflections in a stereotyped back and forth pattern that caused them to oscillate around the balance point and reduce the frequency of crashes. To quantify this across subjects, we used a metric motivated by Park et al.,[14] calculating the standard deviation of the phase difference between angular position and joystick deflections (STD_{Joy_Pos}). The Proficient group learned to statistically increase STD_{Joy_Pos}, meaning that they used more varied joystick deflections in relation to the MARS angular position. By contrast, the Not-Proficient group was unable to do this and had a value that was statistically distinct and 1.7 times less than the Proficient group. This pattern means that the Not-Proficient Group maintains a stereotyped phase relationship between joystick deflection and MARS angular position.

Why were subjects in the Proficient group able to successfully stabilize around the balance point whereas the Somewhat-Proficient and Not-Proficient groups could not? One possibility is that subjects in the Proficient group have lower vestibular thresholds for angular acceleration and could make better estimates about angular position by integrating angular acceleration signals. Some subjects in the Not-Proficient group remarked that they had difficulty determining when the MARS was moving, suggesting that they had poor vestibular thresholds and may have used the strategy of large sways at high velocities to obtain a better sense of their angular direction. Work by Rosenberg et al.,[15] for example, has shown a correlation between vestibular thresholds and acuity in a manual tracking task.

However, our paradigm has the additional facet of being a skilled motor learning task. One additional possibility is that during skilled motor learning subjects explore a complex solution space that can have many solutions, some of which are suboptimal.[7,11] Poor performers may actually acquire a suboptimal strategy because they do not explore the full solution space and because their strategy may provide a sense of stability,[1] such as reducing the frequency of Crashes at the cost of very large

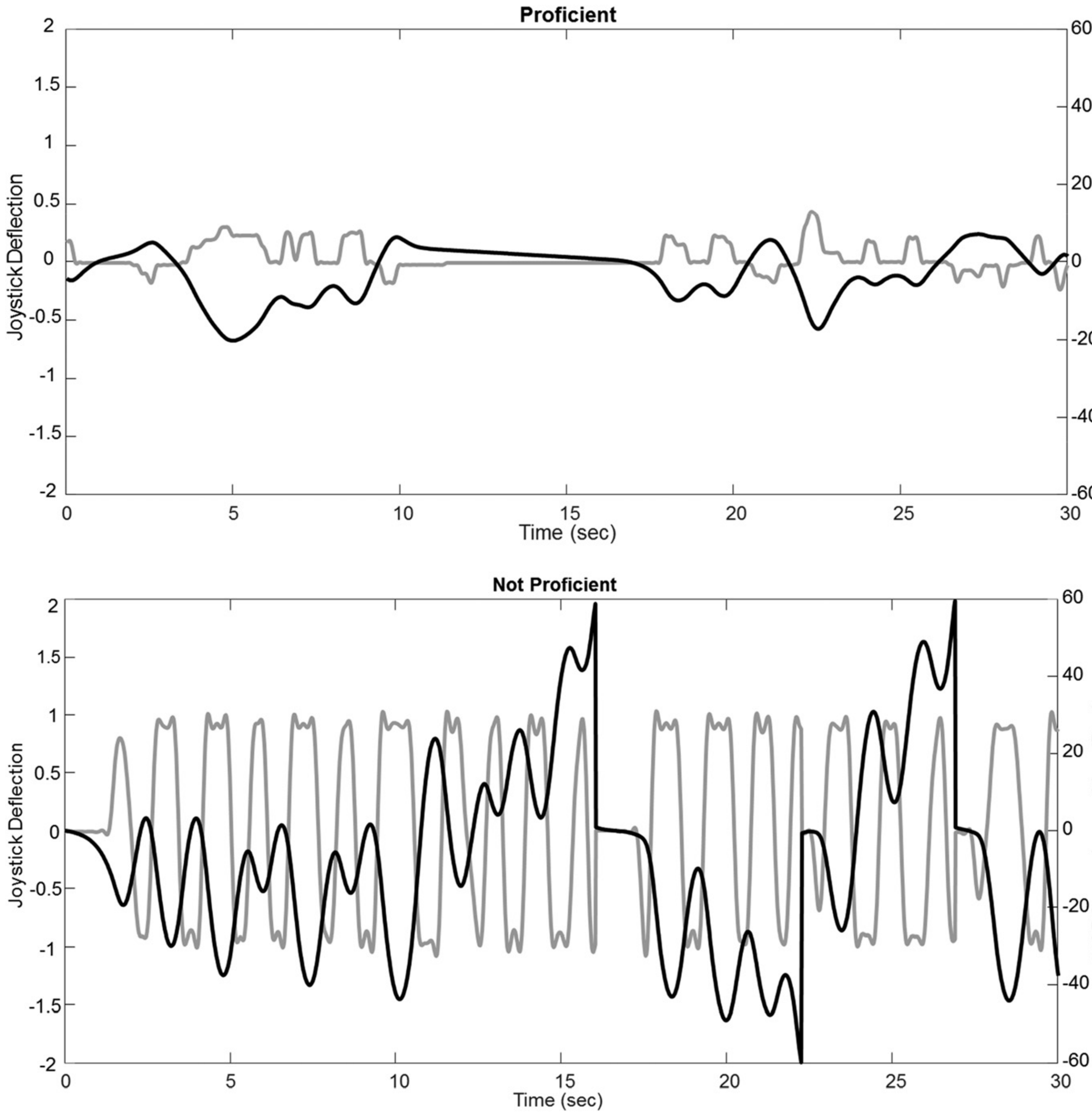

Fig. 5. Trial data of a representative subject from the Proficient Group (top) and from the Not-Proficient group (bottom) for their final trial on Day 2. Joystick deflections (grey line) from the Not-Proficient subject are stereotypical and those from the Proficient subject are not. Black lines are the MARS angular position.

oscillations. Is there a way to ensure that subjects in our task avoid the suboptimal strategy of very stereotyped large magnitude joystick deflections? Ganesh et al.[7] designed a task with multiple solutions and found that even when subjects were shown the best solution they nevertheless tended to use their initial suboptimal strategies. Similarly, in Vimal et al.[18] one group of subjects in the Horizontal Roll condition were given explicit verbal instructions on an optimal strategy of making intermittent small magnitude joystick deflections. This group was statistically no different than the group that received no instructions. Another group underwent a specialized training program that taught them how to dissociate position and motion cues in addition to an optimal joystick strategy.[18] All of those subjects showed significant learning of the optimal strategy, and full retention 4 months later. Together these findings suggest that poor performers may initially have adopted a suboptimal strategy that can be identified using machine learning techniques, and that training programs can be implemented to lead them toward the optimal strategy and performance similar to that of the Proficient group.

ACKNOWLEDGMENTS

This study was supported through a Synergy Grant from the Translational Research Institute for Space Health through NASA NNX16AO69A. The MARS device was developed under Air Force Office of Scientific Research AFOSR FA9550-12-1-0395. VPV and HZ contributed equally to the paper and are co-first authors.

Authors and affiliations: Vivekanand Pandey Vimal, Ph.D., Lila N. Fakharzadeh, M.S., James R. Lackner, Ph.D., and Paul DiZio, Ph.D., Ashton Graybiel Spatial Orientation Laboratory; Dr. Vimal, Dr. Pengyu Hong, Ms. Fakharzadeh, Dr. Lackner, and Dr. DiZio, Volen Center for Complex Systems; Drs. Lackner and DiZio, Department of Psychology; Han Zheng and Dr. Hong, Department of Computer Science, Brandeis University, Waltham. MA.

REFERENCES

1. Babič J, Oztop E, Kawato M. Human motor adaptation in whole body motion. Sci Rep. 2016; 6(1):32868.
2. Blei DM, Jordan MI. Variational inference for Dirichlet process mixtures. Bayesian Anal. 2006; 1(1):121–143.
3. Collins J J, De Luca C. Upright, correlated random walks: A statistical-biomechanics approach to the human postural control system. Chaos. 1995; 5(1):57–63.
4. Collins JJ, De Luca CJ. Open-loop and closed-loop control of posture: a random-walk analysis of center-of-pressure trajectories. Exp Brain Res. 1993; 95(2):308–318.
5. Friedman J. Greedy function approximation: a gradient boosting machine. Ann Statist. 2001; 29(5):1189–1232.
6. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; 33(1):1.
7. Ganesh G, Haruno M, Kawato M, Burdet E. Motor memory and local minimization of error and effort, not global optimization, determine motor behavior. J Neurophysiol. 2010; 104(1):382–390.
8. Graybiel A, Wood CD, Miller EF, Cramer DB. Diagnostic criteria for grading the severity of acute motion sickness Aerosp Med. 1968;39(5):453–455.
9. Lambert JD. Computational methods in ordinary differential equations. introductory mathematics for scientists and engineers. Hoboken (NJ): Wiley; 1973.
10. Mahalanobis PC, editor. On the generalized distance in statistics. National Institute of Science of India; 1936.
11. Morasso P, Casadio M, De Santis D, Nomura T, Rea F, Zenzeri J. Stabilization strategies for unstable dynamics. J Electromyogr Kinesiol. 2014; 24(6):803–814.
12. Panic AS, Panic H, DiZio P, Lackner JR. Gravitational and somatosensory influences on control and perception of roll balance. Aerosp Med Hum Perform. 2017; 88(11):993–999.
13. Panic H, Panic AS, DiZio P, Lackner JR. Direction of balance and perception of the upright are perceptually dissociable. J Neurophysiol. 2015; 113(10):3600–3609.
14. Park S-W, Dijkstra T, Sternad D. Learning to never forget—time scales and specificity of long-term memory of a motor skill. Front Comput Neurosci. 2013; 7:111.
15. Rosenberg MJ, Galvan-Garza RC, Clark TK, Sherwood DP, Young LR, Karmali F. Human manual control precision depends on vestibular sensory precision and gravitational magnitude. J Neurophysiol. 2018; 120(6):3187–3197.
16. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987; 20:53–65.
17. Sammut C, Webb GI,eds. Leave-one-out cross-validation. Encyclopedia of machine learning. Boston (MA): Springer; 2010:600–601.
18. Vimal VP, DiZio P, Lackner JR. Learning and long-term retention of dynamic self-stabilization skills. Exp Brain Res. 2019; 237(11):2775–2787.
19. Vimal VP, DiZio P, Lackner JR. Learning dynamic balancing in the roll plane with and without gravitational cues. Exp Brain Res. 2017; 235:3495–3503.
20. Vimal VP, Lackner JR, DiZio P. Learning dynamic control of body roll orientation. Exp Brain Res. 2016; 234(2):483–492.
21. Vimal VP, Lackner JR, DiZio P. Learning dynamic control of body yaw orientation. Exp Brain Res. 2018; 236(5):1321–1330.