**RESEARCH ARTICLE** | *Control of Movement*

# Interactions between motor exploration and reinforcement learning

**Shintaro Uehara,[1,2] Firas Mawase,[1] Amanda S. Therrien,[3,4] Kendra M. Cherry-Allen,[1] and Pablo Celnik[1,3]**

[1]*Department of Physical Medicine and Rehabilitation, Johns Hopkins Medical Institutions, Baltimore, Maryland;* [2]*Japan Society for the Promotion of Science, Tokyo, Japan;* [3]*Department of Neuroscience, Johns Hopkins Medical Institutions, Baltimore, Maryland; and* [4]*Center for Movement Studies, The Kennedy Krieger Institute, Baltimore, Maryland*

Submitted 12 June 2018; accepted in final form 19 June 2019

**Uehara S, Mawase F, Therrien AS, Cherry-Allen KM, Celnik P.** Interactions between motor exploration and reinforcement learning. *J Neurophysiol* 122: 797–808, 2019. First published June 26, 2019; doi:10.1152/jn.00390.2018.—Motor exploration, a trial-and-error process in search for better motor outcomes, is known to serve a critical role in motor learning. This is particularly relevant during reinforcement learning, where actions leading to a successful outcome are reinforced while unsuccessful actions are avoided. Although early on motor exploration is beneficial to finding the correct solution, maintaining high levels of exploration later in the learning process might be deleterious. Whether and how the level of exploration changes over the course of reinforcement learning, however, remains poorly understood. Here we evaluated temporal changes in motor exploration while healthy participants learned a reinforcement-based motor task. We defined exploration as the magnitude of trial-to-trial change in movements as a function of whether the preceding trial resulted in success or failure. Participants were required to find the optimal finger-pointing direction using binary feedback of success or failure. We found that the magnitude of exploration gradually increased over time when participants were learning the task. Conversely, exploration remained low in participants who were unable to correctly adjust their pointing direction. Interestingly, exploration remained elevated when participants underwent a second training session, which was associated with faster relearning. These results indicate that the motor system may flexibly upregulate the extent of exploration during reinforcement learning as if acquiring a specific strategy to facilitate subsequent learning. Also, our findings showed that exploration affects reinforcement learning and vice versa, indicating an interactive relationship between them. Reinforcement-based tasks could be used as primers to increase exploratory behavior leading to more efficient subsequent learning.

**NEW & NOTEWORTHY** Motor exploration, the ability to search for the correct actions, is critical to learning motor skills. Despite this, whether and how the level of exploration changes over the course of training remains poorly understood. We showed that exploration increased and remained high throughout training of a reinforcement-based motor task. Interestingly, elevated exploration persisted and facilitated subsequent learning. These results suggest that the motor system upregulates exploration as if learning a strategy to facilitate subsequent learning.

meta-learning; motor exploration; reinforcement learning; savings; trial and error

## INTRODUCTION

When learning new motor behaviors, such as when trying a new sport or relearning daily activities after neurological injury, trial and error plays a pivotal role. This process, known as motor exploration, helps the motor system to identify the consequences of various actions and update their values based on each movement outcome. This, in turn, allows the person to regulate the expression of the probed actions (Dhawale et al. 2017). Motor exploration has been linked to reinforcement learning (Dhawale et al. 2017), where actions leading to favorable outcomes (i.e., reward) are reinforced and become more frequently expressed while those resulting in unfavorable outcomes are avoided (Sutton and Barto 1998).

Previous animal and human studies have investigated the impact of motor exploration on motor learning outcomes. Courtship vocalization studies in songbirds indicate that rendition-to-rendition variability in the pitch of their vocalizations, thought to partly serve as motor exploration, supports continuous learning and further optimization of vocalization performance (Fiete et al. 2007; Tumer and Brainard 2007). Similarly, human behavioral studies demonstrated that the magnitude of movement variability or motor exploration is associated with learning rate in a reinforcement-based arm-reaching task that required movement modification toward an optimal pattern (Chen et al. 2017; Therrien et al. 2016; Wu et al. 2014).

Although exploration is critical for reinforcement learning, it is not completely understood whether and how the amount of motor exploration changes over the course of reinforcement learning. Learning-related modification in motor exploration and its underlying neural circuit mechanisms were investigated in songbird studies. Here, as learning proceeds exploratory variability in song production decreases, a change associated with shifts in the control of motor program away from the forebrain to descending motor pathways for vocal control (Aronov et al. 2008) via synaptic reorganization (Garst-Orozco et al. 2014). Much less is known, however, about the interactions between exploration and reinforcement learning in humans. Considering the contribution of exploration in reinforcement learning, it is conceivable that the amount of exploratory behavior is elevated in the early stages of learning, when action values are still unknown, but then reduces as one finds a motor pattern that more reliably leads to reward (i.e., optimal motor solution). Alternatively, it is possible that the amount of motor exploration remains elevated throughout

Address for reprint requests and other correspondence: P. Celnik, Dept. of Physical Medicine and Rehabilitation, Johns Hopkins Medical Institutions, 600 N. Wolfe St., Baltimore, MD 21287 (e-mail: pcelnik@jhmi.edu).

training. This could occur if the motor system not only learns the specific task but also acquires knowledge to be exploratory, allowing it to quickly find an optimal solution when exposed to similar conditions in the future (e.g., learning to learn; Braun et al. 2009, 2010; Krakauer and Mazzoni 2011).

In the present study, we investigated temporal changes in the amount of motor exploration while healthy participants trained on a goal-directed finger-pointing task. The task was designed to follow a reinforcement learning paradigm in which participants were required to adjust their pointing movement in a trial-by-trial manner toward a predetermined target direction (unbeknown to participants), relying solely on binary feedback about performance outcome (Izawa and Shadmehr 2011; Pekny et al. 2015; Therrien et al. 2016; Uehara et al. 2018). We defined exploration as the magnitude of trial-to-trial change in movement direction as a function of whether the preceding trial resulted in success or failure. We first studied whether and how the magnitude of motor exploration changes over the course of training on the task (*experiment 1*). Second, we assessed the impact that modulation of motor exploration may have on subsequent training on the same task (*experiment 2*) or on a task that requires a different motor solution (*experiment 3*). Finally, we reassessed how the magnitude of motor exploration changes over the course of longer task training (*experiment 4*).

## MATERIALS AND METHODS

### *Participants*

The study was reviewed and approved by the Johns Hopkins University School of Medicine Institutional Review Board and was in accordance with the Declaration of Helsinki. A total of 59 healthy participants {25.6 ± 6.3 yr [mean ± standard deviation (SD)]; 35 women and 24 men; 53 right-handers (self-reported)} were recruited for the study. All individuals were naive to the purpose of the study. They provided written informed consent before participating in the study. None of the participants had a history of neurological disease and/or psychological disorders.

### *Finger-Pointing Task*

Participants performed a center-out finger-pointing task, moving a visually displayed cursor from a central starting location through a target in a shooting movement (Fig. 1*A*). Participants sat ~45 cm in front of a vertical computer monitor (1,280 × 1,024-pixel resolution). They were instructed to move a digitizing stylus attached on the ventral surface of the index finger on their dominant hand over a horizontal digitizing tablet (48.8 × 30.5-cm active area, Intuos4 XL; Wacom, Saitama, Japan) located on a table. Thus participants mainly moved the metacarpophalangeal joint of the index finger to control the stylus movement. To facilitate the motions while relaxing the hand, we asked participants to rest their forearm on an arm support sling that allowed free movements of the arm while eliminating the need for gravitational support. The tablet and participants' forearm were covered by a box to prevent participants from directly looking at their hand while performing the task. The position of the stylus, sampled at 60 Hz through a custom MATLAB program (R2015b; MathWorks), corresponded to the position of a yellow 1.5-mm-diameter cursor displayed on a black screen such that moving the stylus forward moved the cursor upward. The mapping between the stylus and the displayed cursor displacement (mm) was set as 1:2.

In the task, participants attempted to move the displayed cursor rapidly from a white 3-mm-square starting position centered in the middle of the screen toward a white 16-mm-diameter target in a

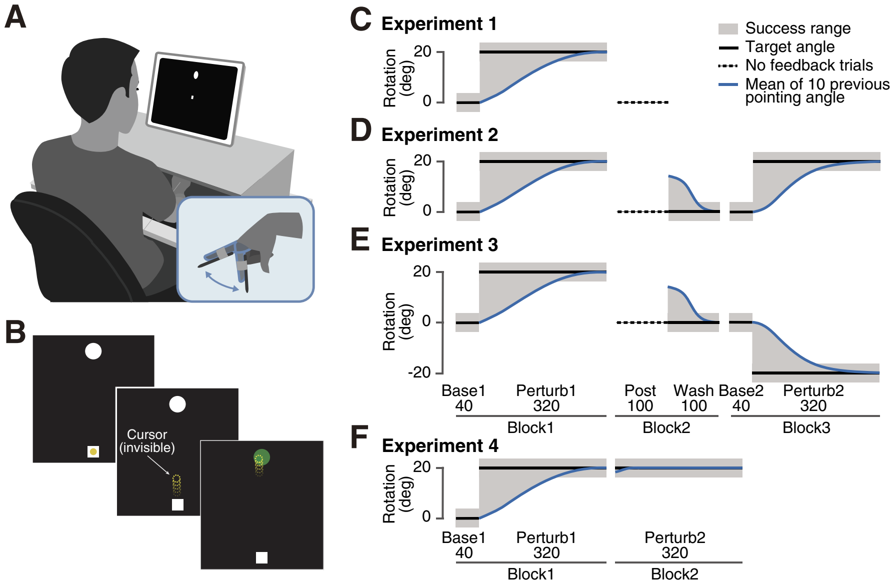

Fig. 1. Experimental protocols. *A*: experimental setup. Participants moved a stylus on a digitizing tablet with their index finger. *B*: finger-pointing task. Participants performed a pointing movement from a central starting position to around a visible white target. Binary feedback (target color) about task performance was presented instead of vector cursor feedback. Participants were instructed to obtain positive (green) color feedback at every trial. *C*: *experiment 1*. Participants performed the task in 2 blocks, composed of baseline (Base1), perturbation (Perturb1), and postperturbation (Post) phases. During the Perturb1 phase, a range for success feedback (gray area) gradually shifted from the original range toward counterclockwise direction according to a moving average of the previous 10 pointing directions (blue line). No-feedback trials were implemented for the Post phase. *D*: *experiment 2*. Participants performed the task in 3 blocks, composed of Base1, Perturb1, Post, washout (Wash), Base2, and Perturb2 phases. During the Wash phase, the binary feedback was presented so that pointing direction returned to the baseline level. The task setting in the following Base2 and Perturb2 phases was exactly the same as in Base1 and Perturb1. *E*: *experiment 3*. Participants performed the task in a setting similar to *experiment 2*. The only difference was that the target angle was set at clockwise direction during the Perturb2 phase. Numbers under each phase in the *x*-axis represent the number of trials. The *y*-axis represents rotation angle from the center of the visible target; positive value indicates counterclockwise rotation. Note that the direction of the perturbation was set to the opposite side for left-handed participants. *F*: *experiment 4*. Participants performed the task in a setting similar to *experiment 1* but with more trials for the Perturb phase.

straight line with no corrections (Fig. 1*B*). The visible target was always displayed at 90°, 10 cm superior to the starting position. The trials began when the cursor was held in the starting position for 500 ms, leading to the appearance of the target on the screen. Upon presentation of the target, participants started to move their finger so that the cursor crossed through the target. However, the cursor disappeared immediately after participants moved out of the starting position (>0.3 mm). In other words, participants did not receive online cursor feedback. Instead, reinforcing binary color feedback (green: success, red: failure) was presented to participants at the moment when the invisible cursor passed through the 10-cm-radius boundary circle centered around the starting position (i.e., the movement end point). The target's color turned green if the end point was within an invisible "success range" (e.g., between the target's bounds) or red if it missed the range. If a participant's movement was too fast (<100 ms) or too slow (>300 ms), a high- or low-pitched auditory tone was provided. Thus the task was designed so that the movements were not ballistic but constrained to be executed within a predefined time period. We verbally instructed participants to prioritize task success and then, if possible, to complete the movement in the allowed time window (Table 1). After each trial, participants moved back to the starting position, guided by a yellow ring that indicated the distance of the current cursor position from the central starting position. When the invisible cursor was within 1.25 cm from the central starting position, the ring was transformed into the visible cursor.

### *Experimental Procedure*

*Experiment 1.* We examined whether and how the magnitude of motor exploration changed during training on the reinforcement-based learning task. Twenty participants (25.1 ± 5.7 yr; 14 women and 6 men) performed the finger-pointing task in two consecutive blocks (Fig. 1*C*). The first block started with a 40-trial (1 epoch) "baseline" phase (Base1) in which the invisible success range was kept constant between the visible target's bounds (90 ± 4.5° on the screen). This was followed by a 320-trial (40 trials × 8 epochs) "perturbation" phase (Perturb1) in which the success range was gradually perturbed from the original range, unbeknown to participants. This was intended to have participants gradually learn a new pointing direction toward a predetermined target angle, set at 110° on the screen (20° counterclockwise rotation from the visible target) through a trial-and-error process (Therrien et al. 2016; Uehara et al. 2018). For this purpose, the left bound of the success range was rotated to 114.5°, whereas the right bound gradually shifted in a trial-by-trial manner according to a moving average of 10 previous pointing directions (i.e., in a "closed-loop" reinforcement schedule). Therefore, to obtain positive feedback, participants were basically required to adjust their pointing direction toward counterclockwise rotation beyond the average of their last 10 trials. When the moving average of 10 last trials fell within the range of 105.5° to 114.5° (the target angle ±4.5°), this range became the success range so that pointing direction could stay around the target angle. If the moving average exceeded 114.5°, the success range was set between 105.5° and the angle of the moving average to lead pointing direction back to the target angle.

Table 1. *Movement time*

| | *Block 1* | *Block 2* | *Block 3* |
|---|---|---|---|
| *Experiment 1* | 251.8 ± 63.3<br>(18.3 ± 8.5%) | 198.9 ± 40.2<br>(12.5 ± 7.6%) | |
| *Experiment 2* | 222.9 ± 35.8<br>(13.1 ± 6.3%) | 197.5 ± 28.8<br>(9.8 ± 6.3%) | 191.6 ± 36.0<br>(11.2 ± 6.9%) |
| *Experiment 3* | 257.5 ± 52.6<br>(15.4 ± 4.3%) | 214.9 ± 39.1<br>(9.2 ± 5.4%) | 207.7 ± 51.0<br>(11.1 ± 6.5%) |
| *Experiment 4* | 224.2 ± 25.1<br>(10.9 ± 3.3%) | 199.6 ± 31.6<br>(7.6 ± 4.8%) | |
| Nonlearners | 264.4 ± 54.8<br>(18.8 ± 7.2%) | 221.2 ± 61.7<br>(12.1 ± 7.4%) | |

Values indicate the average (± SD) movement time in milliseconds and % of number of trials outside the time window (in parentheses) in each block. Note that participants included in *experiments 1–3* are only those classified as learners.

After taking a break of a few minutes, the participants proceeded to the second block consisting of a 100-trial "postperturbation" phase (Post). During this phase, the target always turned black irrespective of pointing angles (i.e., no-feedback trials). Before starting the Post phase, we displayed written instructions to participants to repeat the same movements as previously done to make the target green. We implemented these trials to investigate whether adjusted pointing movements continued under the condition in which no factor encouraging further motor adjustments was presented. To ensure that participants maintained their motivation, we presented the percentage of successful trials on the screen at the end of the first block.

Note that the perturbed direction during the Perturb1 phase was set opposite for left-handers, guiding their pointing direction converged around 70° on the screen (20° clockwise rotation from the visible target). For simplicity's sake, all other protocols are presented and illustrated in a setting for right-hand-dominant individuals.

*Experiment 2.* Results in *experiment 1* showed that the magnitude of motor exploration was elevated and remained high throughout the Perturb1 phase. One possible explanation for this finding is that elevated exploration may persist in order to improve the efficiency of learning during subsequent exposures to the same training situation. To test this assumption, we recruited 15 new participants (27.3 ± 6.4 yr; 10 women and 5 men) and asked them to perform the task in three consecutive task blocks (Fig. 1*D*). The first block was composed of the 40-trial Base1 phase and the 320-trial Perturb1 phase as in *experiment 1*. In the second block, the 100-trial Post phase occurred first, followed by a 100-trial "washout" (Wash) phase. During the Wash phase, the reinforcing binary feedback was presented again so that the participants were able to gradually get their pointing direction back to the baseline level (toward the visible target) under the same learning context as in the Perturb1 phase. For this purpose, the target angle was set at 90° and the success range was adjusted in a trial-by-trial manner on the basis of the closed-loop reinforcement schedule. The third block was composed of a 40-trial second baseline phase (Base2) and a 320-trial second perturbation phase (Perturb2). The perturbation rule for this block was in line with that implemented in the first block. The percentage of successful trials was presented to the participants at the end of each block. A few minutes of breaks were inserted between the blocks.

*Experiment 3.* Results in *experiment 2* showed that participants were able to relearn the same task more quickly in the second training session, in association with increased exploration from the beginning of the second exposure. It is possible, however, that the rapid relearning was not due to increased exploration but rather to the presence of a directional bias toward the same rotation direction needed in the second training session. Indeed, previous studies suggested that directional bias in movement and/or in cortical motor representation can be formed via repetition of successful movements (Diedrichsen et al. 2010; Huang et al. 2011; Mawase et al. 2017b; Verstynen and Sabes 2011). To rule out this confounder, we performed an additional experiment in which participants were required to adjust their pointing movement during the second training session in the opposite direction to that made in the first training. We recruited a new group of 14 participants (27.8 ± 7.8 yr; 8 women and 6 men) and asked them to perform a task in three consecutive blocks (Fig. 1*E*). The task setting in the first two blocks was exactly the same as in *experiment 2*. The only difference in this experiment was in the third block, where the target angle was set at 70° on the screen (20° clockwise rotation from the visible target) during the Perturb2 phase.

*Experiment 4.* In the previous three experiments, we found that the magnitude of motor exploration was gradually elevated and stayed high throughout the first training. Thus we asked whether the increased exploratory behavior changes when participants repeat the same tasks for longer periods of time. To investigate this, we recruited a new group of 10 participants (21 ± 0.0 yr; 3 women and 7 men) and asked them to perform the same task as in *experiment 1* but with a double number of trials (Fig. 1*F*). The first block was composed of the 40-trial Base1 phase, followed by the 320-trial Perturb1 phase and then the second block of another 320-trial Perturb2 phase. The participants were allowed to have a 2- to 3-min break between the blocks to prevent fatigue effects.

### *Data Analysis*

*Task performance.* Task performance was quantified with pointing angle (PA), the angle between the line connecting the starting position to the center of the visible target and the line connecting the starting position to the end point. To analyze right- and left-handed dominances together, we flipped the left-handed data to correspond to right-handed participants. We excluded trials in which PA exceeded $|60°|$ as outliers ($<0.5\%$ of trials among all the participants). To quantify learning rate during training sessions, we measured "initial deviation" in PA as follows: initial deviation (ID) = $\text{PA}_{\text{Perturb}} - \text{PA}_{\text{Base}}$, where $\text{PA}_{\text{Perturb}}$ represents the mean PA of the first epoch in the perturbation phases (*trials 41–80*) and $\text{PA}_{\text{Base}}$ represents the mean PA in the preceding baseline phase (*trials 1–40*). For data in *experiment 3* only, we flipped the sign of the individual initial deviation from the second training session to match that of the first training session. We also assessed the training-related performance in terms of the mean percentage of successful trials for each of 40-trial epochs on which binary feedback was presented.

*Motor exploration.* For participants to learn the task, they needed to actively explore new possible movement directions. Especially when the previous trial resulted in a negative outcome, they needed to change movements in the search for a better pointing direction that more reliably led to a positive outcome. Therefore, we defined the magnitude of trial-to-trial change in PA after failed trials as a measure of the amount of motor exploration (Pekny et al. 2015; Sidarta et al. 2016). In addition, since we wanted to capture the true magnitude changes of the movement deviation regardless of direction, we computed the unsigned magnitude of PA changes ($|\Delta\text{PA}|$) from trial $n$ to trial $n + 1$ contingent upon trial $n$ being a success ($S = 1$) or a failure ($S = 0$). We then calculated the mean of $|\Delta\text{PA}|$ after successful or after failed trials for each of 40-trial epochs to track changes in the amount of motor exploration over the course of task training.

### *Statistical Analysis*

To test whether and how the magnitude of motor exploration changes during task training, we performed a repeated-measures analysis of variance (RM ANOVA) on $|\Delta\text{PA}|$ across epochs during the perturbation phases. For *experiment 1*, a two-way RM ANOVA was performed with within-participant factors of Outcome (success, failure) and Epoch (8 epochs). For *experiments 2* and *3*, a three-way RM ANOVA was performed with within-participant factors of Outcome, Epoch, and Session (1st and 2nd training sessions).

For *experiments 2* and *3* only, we separately performed a paired *t*-test between $|\Delta\text{PA}|$ of the first epoch in the Perturb1 phase and that in the Perturb2 phase to evaluate the a priori hypothesis that exploratory behavior would be greater from the beginning of the second than the first training session. A paired *t*-test was also used to compare initial deviation in PA (a proxy for learning rate) between the first and second training sessions. To assess the relationship between the amount of exploration at the beginning of task training and learning rate, we applied Pearson's correlation analysis between $|\Delta\text{PA}|$ of the first epoch in the perturbation phase and initial deviation. This was separately performed for the first and second training sessions. Finally, to test whether the changes in the magnitude of exploration from the first to the second training sessions were associated with changes in learning rate between the two sessions, we performed Pearson's correlation analysis between ($|\Delta\text{PA}|_{\text{Second}} - |\Delta\text{PA}|_{\text{First}}$) and ($\text{ID}_{\text{Second}} - \text{ID}_{\text{First}}$), where $|\Delta\text{PA}|_{\text{First}}$ and $|\Delta\text{PA}|_{\text{Second}}$ represent $|\Delta\text{PA}|$ of the first epoch in the Perturb1 and Perturb2 phases and $\text{ID}_{\text{First}}$ and $\text{ID}_{\text{Second}}$ represent initial deviation in PA in the first and second training sessions, respectively.

To determine whether participants maintained the learned pointing direction during the Post phase, we compared the average PA of the first 40-trial epoch in the Post phase to the epoch of the Base1 phase with a paired *t*-test. Similarly, to determine whether pointing direction successfully returned to baseline level during the Wash phase and the subsequent Base2 phases, we compared the epoch in the Base1 phase to the last 40-trial epoch in the Wash phase and the epoch in the Base2 phase, respectively, with a paired *t*-test.

All statistical analyses were performed with SPSS (version 20; IBM, Armonk, NY). All RM ANOVAs were tested for the assumption of homogeneity of variance with Mauchly's test of sphericity. For those tests in which this assumption was violated, the Greenhouse-Geisser correction statistic was reported. Effects were considered significant if $P \leq 0.05$. Effect sizes were reported in Cohen's $d_z$ value for paired *t*-test, Cohen's $d$ value for unpaired *t*-test, and partial eta squared value ($\eta_p^2$) for ANOVA.

## RESULTS

### *Classification of "Learners" and "Nonlearners"*

As observed in our previous study (Uehara et al. 2018), we found a subset of participants who did not adjust their PA sufficiently to reach the target angle in a limited number of perturbed trials. These nonlearners were analyzed separately from the rest of the participants (learners). Moreover, they served as a control group for our investigation into the relationship between a modification in motor exploration and task learning, rather than simple task execution. We defined participants as nonlearners if the moving average of PAs did not exceed $|18°|$ (90% of the target angle) during the Perturb1 phase (320 trials). With this criterion, 5 of 20 (*experiment 1*), 3 of 15 (*experiment 2*), and 2 of 14 participants (*experiment 3*) were classified as nonlearners (~20% of total participants but no participants from *experiment 4*), and all the nonlearners were integrated into one group (Nonlearner group; $n = 10$, 6 women and 4 men; 26.7 ± 7.1 yr). Consequently, 15 (23.9 ± 3.7 yr; 11 women and 4 men), 12 (29.0 ± 6.9 yr; 9 women and 3 men), and 11 (26.7 ± 7.1 yr; 6 women and 5 men) participants from *experiments 1*, *2*, and *3*, respectively, were defined as learners and proceeded to the main analyses. Note that one participant from *experiment 3* was excluded as an outlier from the analysis since she/he did not show gradual and systematic learning of the task. Specifically, this participant's moving average of 10 PAs moved back to the baseline level even after reaching the predetermined target angle once (20° counterclockwise rotation) during the Perturb1 phase. Furthermore, the participant moved into the clockwise rotation direction beyond the baseline angle.

### *Motor Exploration Increased During Training*

In *experiment 1*, we investigated whether and how the amount of exploratory behavior changes during training on the reinforcement-based motor task.

We first confirmed that the participants successfully learned the task, indicated by PA shifts on trial-by-trial bases from the original toward the target direction while exposed to the perturbation (Fig. 2). During the Base1 phase, the participants performed the task accurately given that pointing direction was within the visible target [mean PA −1.0 ± 0.7° (mean ± SE); Fig. 2*A*] with high accuracy (% of trial success 70.2 ± 4.3%; Fig. 2*B*). During the Perturb1 phase, PA gradually shifted toward the predetermined target angle and finally converged on 18.9 ± 0.8° in the last 40-trial epoch (Fig. 2*A*). This resulted in significant errors early on, accompanied by a slight but gradual increase in the percentage of successful trials across epochs (Fig. 2*B*). This new movement direction was maintained during the subsequent Post phase (average of the first 40 trials 13.4 ± 1.9°) compared with the Base1 phase (paired *t*-test, $t_{14} = 7.2$, $P < 0.001$, $d_z = 1.85$), indicating retention of the newly learned motor pattern.

As a proxy for motor exploration, we computed the magnitude of trial-to-trial changes in PA (|ΔPA|) as a function of whether the initial trial resulted in success or failure. We found that on average |ΔPA| after failed trials was greater than that after successful trials {RM ANOVA [Outcome (2) × Epoch (8)], effect for the factor Outcome $F_{1,\ 14} = 52.4$, $P < 0.001$, $\eta_p^2 = 0.79$; Fig. 2*C*}. This was also revealed by broader probability distributions of |ΔPA| for trials after failure than after success (Fig. 2*D*, *top*). These results indicate that failing to get positive feedback led to greater trial-to-trial movement changes, presumably in search for a better motor solution (Pekny et al. 2015; Sidarta et al. 2016). Importantly, when we focused on temporal changes in |ΔPA| over the course of training, we found that the magnitude of |ΔPA| after failed trials was not constant but gradually increased across epochs and remained elevated throughout the Perturb1 phase (Fig. 2*C*). Though small, similar temporal change was observed in |ΔPA| after successful trials (effect for the factor Epoch $F_{7,\ 98} = 2.7$, $P = 0.01$, $\eta_p^2 = 0.16$; Outcome × Epoch interaction $F_{7,\ 98} = 1.3$, $P = 0.25$, $\eta_p^2 = 0.09$). This increase was qualitatively represented as a broader probability distribution of |ΔPA| at the last (8th) epoch compared with the first epoch (Fig. 2*D*, *top*). In a similar vein, we found wider distribution in the signed magnitude of trial-to-trial changes (ΔPA; Fig. 2*D*, *bottom*) into both positive and negative directions in the last epoch (standard deviation of ΔPA 9.0 ± 1.0°) relative to the first epoch (5.7 ± 0.9°, paired *t*-test, $t_{15} = 3.0$, $P = 0.01$, $d_z = 0.76$). Although there were small biases of the movement-correcting direction in the counterclockwise direction particularly after failed trials (Fig. 2*D*, *bottom*), the increase in |ΔPA| can be regarded as increased exploratory behavior in both counterclockwise and clockwise directions rather than in one particular direction.

In sum, these results demonstrate that the amount of motor exploration is gradually upregulated and remained elevated throughout training on the task.

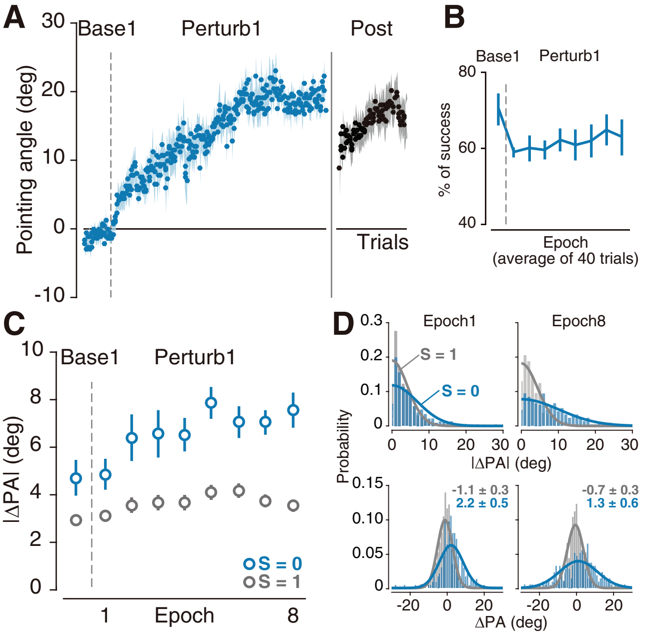

Fig. 2. Gradual increase in exploratory behavior during task training. *A*: pointing angle (PA) in the baseline (Base1), perturbation (Perturb1), and postperturbation (Post) phases. Positive value indicates counterclockwise direction relative to the target position. Dots and shaded areas show the mean and SEs for each trial. Solid vertical line indicates time break between the blocks. *B*: mean % of successful trials for each 40-trial epoch. *C*: mean trial-to-trial unsigned changes in PA (|ΔPA|) for each epoch. Blue and gray open dots indicate trials after failure ($S = 0$) and success ($S = 1$), respectively. *D*: probability distribution of unsigned (|ΔPA|, *top*) and signed (ΔPA, *bottom*) trial-to-trial changes in PA ($S = 0$, $S = 1$) from the first (*left*) and last (*right*) epochs during the Perturb1 phase. The fits of unsigned or signed changes to a folded normal or a normal distribution are plotted over the histogram. Data from all the participants are pooled together on this analysis. Values inside *bottom* panels show the mean and SE of |ΔPA| across participants.

### *Increased Exploration Continued and Facilitated Relearning of the Same Task*

Given that the amount of motor exploration increased and persisted throughout training, in *experiment 2* we tested whether this increase was maintained and benefited subsequent training on the same reinforcement-based task.

First, we replicated the results of *experiment 1* in a separate group of participants; that is, training of the task resulted in a gradual increase in exploratory behavior during the first training exposure (Fig. 3). During the Base1 phase, the participants' pointing direction converged on the visible target (mean PA 0.5 ± 0.9°; Fig. 3*A*) with high accuracy (% of trial success 69.9 ± 4.8%; Fig. 3*B*). When exposed to the Perturb1 phase, the PA gradually shifted (initial deviation 3.3 ± 0.9°; Fig. 3*C*) to end up near the target angle of 18.4 ± 1.0° in the last 40-trial epoch (Fig. 3*A*). This was accompanied by a low number of successful trials early on with a gradual increase across epochs (Fig. 3*B*). The new PA persisted to some degree during the subsequent Post phase (average of first 40 trials 9.8 ± 1.7°), remaining significantly greater compared with Base1 (paired *t*-test, $t_{11} = 5.1$, $P < 0.001$, $d_z = 1.46$). Importantly, as in *experiment 1*, we found gradual increase in |ΔPA| across epochs during the Perturb1 phase along with task learning (Fig. 3, *D* and *E*). This increase was evident after failed trials but not after successful trials.

Second, we found that during the second exposure to the same task the participants were able to adjust their pointing direction more quickly than in the first training session (savings effect). Of note, we confirmed that before the second exposure participants' PA returned to the baseline level during the Wash phase (i.e., learned movement pattern was washed out; average of last 40 trials 2.1 ± 1.0°; paired *t*-test comparing to Base1, $t_{11} = 1.2$, $P = 0.25$, $d_z = 0.35$; Fig. 3*A*). This PA persisted during the following Base2 phase (0.9 ± 0.7°; $t_{11} = 0.3$, $P = 0.79$, $d_z = 0.08$) and resulted in a high percentage of successful

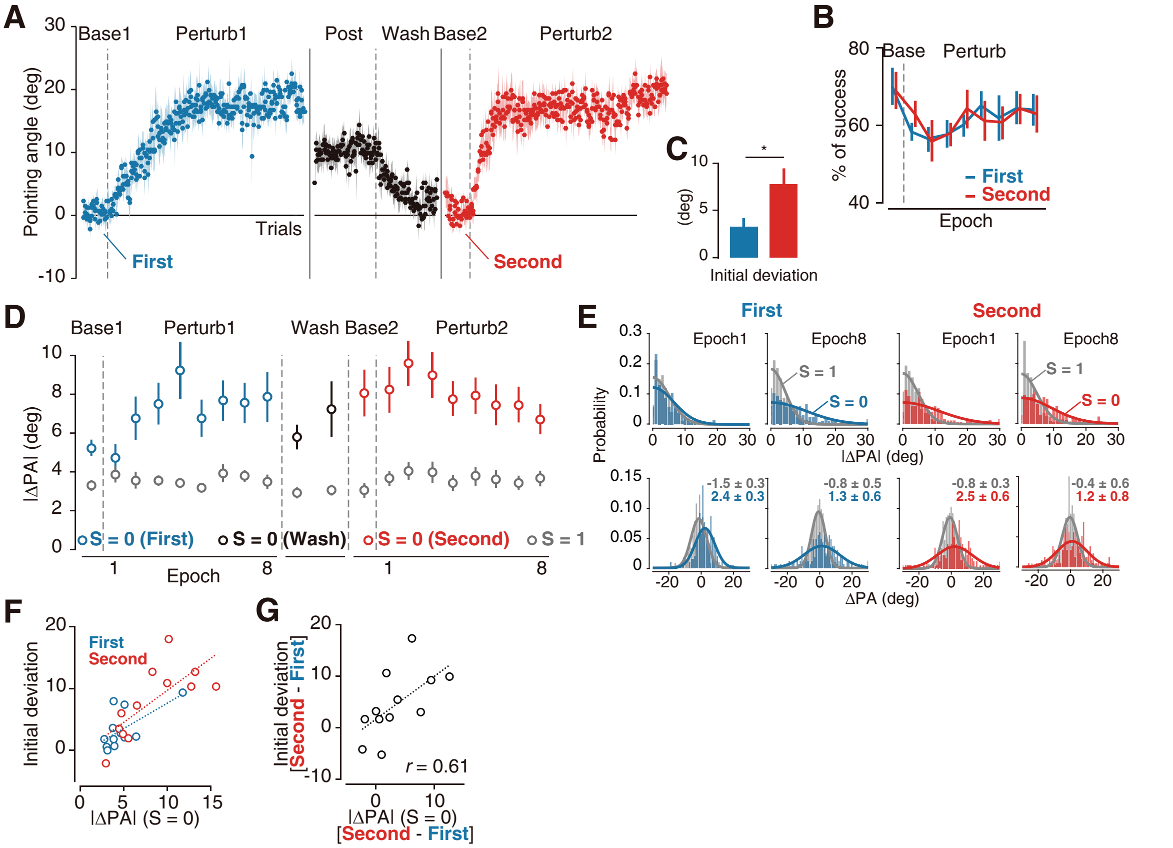

Fig. 3. Persistence of increased exploration and faster relearning during exposure to the second training of the same task. *A*: pointing angle (PA) while training the task. Dots and shaded area show the mean and SE for each trial in the Base1-Perturb1 (blue), Post-Wash (black), and Base2-Perturb2 (red) phases. Solid vertical lines indicate time breaks between the blocks. *B*: mean % of successful trials for each epoch. *C*: initial deviation in PA for the Perturb1 (blue) and Perturb2 (red) phases. $^*P < 0.05$. *D*: unsigned trial-to-trial change in PA (|ΔPA|) across epochs. Blue and red dots indicate trials after failure ($S = 0$) in the 1st and 2nd training sessions, respectively. Gray dots indicate trials after success ($S = 1$). *E*: probability distribution of unsigned (|ΔPA|, *top*) and signed (ΔPA, *bottom*) trial-to-trial changes in PA from the first and last epochs during the Perturb1 and Perturb2 phases. The fits of unsigned or signed changes to a folded normal or a normal distribution are plotted over the histogram. Dashed lines indicate regression line. Values inside *bottom* panels show the mean and SE of ΔPA across participants. *F*: relationship between |ΔPA| ($S = 0$, 1st epoch) and initial deviation for the Perturb1 (blue) and Perturb2 (red) phases. *G*: relationship between the magnitude of changes in |ΔPA| ($S = 0$, 1st epoch) and changes in initial deviation from the 1st to the 2nd training session.

trials (69.0 ± 4.9%; Fig. 3*B*). When participants were exposed to the Perturb2 phase, however, PA shifted toward the target direction more quickly (initial deviation 7.8 ± 1.6°) than during the Perturb1 phase (paired *t*-test, $t_{11} = 2.5$, $P = 0.03$, $d_z = 0.71$; Fig. 3*C*). At the end of training, the participants expressed an amount of angular changes similar to the Perturb1 phase (19.4 ± 1.4° in last 40-trial epoch). Strikingly, |ΔPA| after failed trials remained increased throughout the Wash phase (last 40-trial epoch in Perturb1 7.9 ± 1.3°, Wash 7.3 ± 1.4°; $t_{11} = 0.3$, $P = 0.7$, $d_z = 0.76$). This was followed by greater error corrections from the onset of the Perturb2 phase (1st epoch 8.2 ± 1.2°) compared with that in the Perturb1 phase (4.7 ± 0.7°, paired *t*-test, $t_{11} = 2.6$, $P = 0.02$, $d_z = 0.76$; Fig. 3*D*). Furthermore, this increased magnitude in |ΔPA| after failed trials persisted throughout training, unlike the pattern observed during the first training session {RM ANOVA [Outcome (2) × EPOCH (8) × Session (2)], interaction among 3 factors $F_{7,\ 77} = 3.4$, $P = 0.003$, $\eta_p^2 = 0.24$; Epoch × Session interaction $F_{7,\ 77} = 2.2$, $P = 0.045$, $\eta_p^2 = 0.17$; effect for the factor Outcome $F_{1,\ 11} = 32.6$, $P < 0.001$, $\eta_p^2 = 0.75$}. Additionally, we found no difference in the magnitude of |ΔPA| after successful trials at the training onset between the two sessions (Perturb1 3.9 ± 0.4°, Perturb2 3.7 ± 0.4°; paired *t*-test, $t_{11} = 0.5$, $P = 0.66$, $d_z = 0.13$). These changes in exploratory behavior were qualitatively visualized in a broader probability distribution of both unsigned and signed magnitude of trial-to-trial changes (Fig. 3*E*), albeit small biases of the movement-correcting direction in the direction of the perturbed success range could be observed (Fig. 3*E*, *bottom*).

To determine potential associations between learning rate and the amount of exploration, we performed a correlation analysis. In both the Perturb1 and Perturb2 phases we found a positive correlation between |ΔPA| after failed trials from the first epoch and initial deviation ($r = 0.64$, $P = 0.03$ and $r = 0.74$, $P = 0.01$, respectively; Fig. 3*F*). Furthermore, the magnitude of changes in |ΔPA| after failed trials from the first to the second training session was correlated with the difference in learning rate between the two sessions ($r = 0.61$, $P = 0.04$; Fig. 3*G*). In contrast, we did not find any significant correlations for |ΔPA| after successful trials (Perturb1 $r = 0.35$, $P = 0.26$; Perturb2 $r = 0.10$, $P = 0.76$; changes from Perturb1 to Perturb2 $r = 0.53$, $P = 0.07$).

In sum, these results demonstrate that elevated exploration during the first training session persists during subsequent bouts of training and is associated with facilitation of subsequent learning.

### *Increased Exploration Facilitated Learning on a Second Task That Required a Different Motor Solution*

Although in *experiment 2* increased exploration was associated with learning facilitation during the second training, it is possible that this effect was simply due to the contribution of a directional bias toward the same rotation direction in the second training. To rule out this confounder, a new group of participants took part in *experiment 3*, where the Perturb2 phase required adjusting the pointing direction in the opposite direction to that made in the Perturb1 phase. In this manner, faster relearning would be attributed to increased motor exploration rather than a formed directional bias.

Similar to *experiments 1* and *2*, the amount of motor exploration showed a gradual increase as the participants learned the task during the first training session (Fig. 4). During the first Base1 phase, the participants' pointing direction converged on the visible target (PA 1.5 ± 0.9°; Fig. 4*A*) with high accuracy (% of trial success 66.8 ± 5.2%; Fig. 4*B*). Then, when participants were exposed to the Perturb1 phase, PA gradually shifted (initial deviation 2.6 ± 1.5°; Fig. 4*C*), to end up near the target angle of 17.6 ± 1.1° at the last 40-trial epoch (Fig. 4*A*). This was accompanied by a low number of successful trials early on with a gradual increase across epochs (Fig. 4*B*). During the subsequent Post phase, the new PA still persisted (average of first 40 trials 15.7 ± 2.1°) without returning to the level of the Base1 phase (paired *t*-test, $t_{10} = 5.6$, $P < 0.001$, $d_z = 1.68$). Importantly, we replicated our previous findings that |ΔPA| after failed trials showed a gradual increase across epochs throughout the Perturb1 phase (Fig. 4*D*).

When exposed to the second training session, the participants were able to learn the task more quickly than in the first training, similar to *experiment 2*. We confirmed that the learned PA was washed out, i.e., participants' PA returned to the level of the Base1 phase during the Wash phase (average of last 40 trials 1.5 ± 1.0°; paired *t*-test, $t_{10} = 0.1$, $P = 0.94$, $d_z = 0.02$; Fig. 4*A*). This angle direction persisted during the subsequent Base2 phase (1.6 ± 0.6°; paired *t*-test, $t_{10} = 0.1$, $P = 0.95$, $d_z = 0.02$) and resulted in a high percentage of task success (% of trial success 68.0 ± 6.3%; Fig. 4*B*). Importantly, when exposed to the Perturb2 phase, their PA shifted toward the new target direction (initial deviation 8.8 ± 1.4°) more quickly than during the Perturb1 phase (paired *t*-test, $t_{10} = 4.0$, $P = 0.002$, $d_z = 1.22$; Fig. 4*C*), even when the second training session required adjusting the pointing direction in the opposite direc-

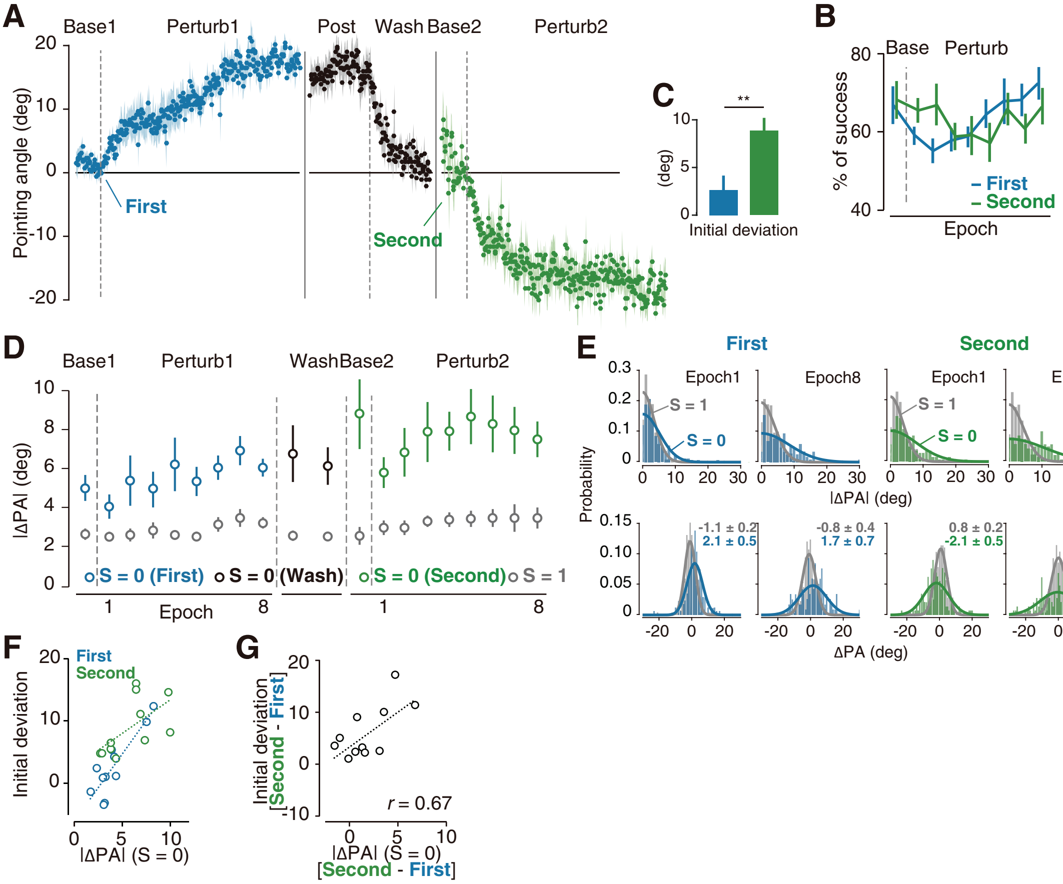

Fig. 4. Increased exploration facilitated learning in the second task requiring a different motor solution. *A*: pointing angle (PA) while training the task. Dots and shaded areas show the means and SE for each trial during the Base1-Perturb1 (blue), Post-Wash (black), and Base2-Perturb2 (green) phases. *B*: mean % of successful trials. *C*: initial deviation in PA during the Perturb1 (blue) and Perturb2 (green) phases. ***P* < 0.01. *D*: unsigned trial-to-trial change in PA (|ΔPA|) across epochs. Blue and green dots indicate trials after failure ($S = 0$) from the first and second training sessions, respectively. Gray dots indicate trials after success ($S = 1$). *E*: probability distribution of |ΔPA| (*top*) and ΔPA (*bottom*) from the first and last epochs during the Perturb1 and Perturb2 phases. Values inside *bottom* panels show the mean and SE of ΔPA across participants. *F*: relationship between |ΔPA| ($S = 0$, 1st epoch) and initial deviation for the Perturb1 (blue) and the Perturb2 (green) phases. *G*: relationship between the magnitude of changes in |ΔPA| ($S = 0$, 1st epoch) and changes in initial deviation from the first to the second training sessions. Dashed lines indicate regression line.

tion (last 40-trial epoch $-18.6 \pm 1.0°$) to that in the first training. Moreover, accompanying this faster learning, |ΔPA| after failed trials remained increased during the Wash phase (last 40-trial epoch in Perturb1 $6.1 \pm 0.4°$, Wash $6.2 \pm 1.0°$; $t_{10} = 0.1$, $P = 0.9$, $d_z = 0.03$) and showed greater magnitude from the onset of the second training compared with the first training (1st epoch of Perturb1 $4.0 \pm 0.6$ °, Perturb2 $5.8 \pm 0.8°$; paired $t$-test, $t_{10} = 2.4$, $P = 0.04$, $d_z = 0.71$; Fig. 4*D*). However, unlike *experiment 2*, |ΔPA| after failed trials continued to increase through the second training {RM ANOVA [Outcome (2) × Epoch (8) × Session (2)], effect for the factor Session $F_{1,\ 10} = 10.4$, $P = 0.009$, $\eta_p^2 = 0.51$}. This temporal pattern was different from that of |ΔPA| after success (Outcome × Session interaction $F_{1,\ 10} = 11.8$, $P = 0.006$, $\eta_p^2 = 0.54$; factor Outcome $F_{1,\ 10} = 41.5$, $P < 0.001$, $\eta_p^2 = 0.81$), in which we found no difference in the magnitude at the onset between the two training sessions (1st epoch of Perturb1 $2.5 \pm 0.2°$, Perturb2 $3.0 \pm 0.3°$; paired $t$-test, $t_{10} = 1.3$, $P = 0.22$, $d_z = 0.40$). Similar to the previous two experiments, changes in exploratory behavior were qualitatively visualized in a broader probability distribution of the magnitude of trial-to-trial changes regardless of small biases of the movement-correcting direction (Fig. 4*E*).

Consistent with *experiment 2* findings, we observed a positive correlation between |ΔPA| after failed trials of the first epoch and initial deviation in both the Perturb1 and Perturb2 phases ($r = 0.87$, $P = 0.001$ and $r = 0.61$, $P = 0.04$, respectively; Fig. 4*F*). Furthermore, changes in |ΔPA| after failed trials from the first to the second training session were correlated with changes in learning rate between the two sessions ($r = 0.67$, $P = 0.03$; Fig. 4*G*). We did not find any significant correlations between |ΔPA| after successful trials and learning rate (Perturb1 $r = 0.45$, $P = 0.16$; Perturb2 $r = 0.36$, $P = 0.27$; changes from Perturb1 to Perturb2 $r = 0.44$, $P = 0.18$).

Although we observed faster learning in the second training session, there still remains a possibility that the second task requiring movement adjustment toward clockwise rotation direction might be by default easier to learn compared with the first task. To rule out this possibility, we reanalyzed data from our previous work (Uehara et al. 2018) in which healthy participants ($n = 12$) performed the same task without a prior training session. The only difference was that in the previous study the target angle was set at 30°, instead of 20°, clockwise rotation from the visible target. The initial deviation in PA showed magnitude ($4.9 \pm 1.2°$) comparable to that of the first task in the present study (unpaired $t$-test, $t_{25} = 0.5$, $P = 0.62$, $d = 0.20$; $t_{22} = 1.1$, $P = 0.28$, $d = 0.45$; $t_{21} = 0.8$, $P = 0.24$, $d = 0.50$, *experiments 1*, *2*, and *3*, respectively). These results indicate that the faster learning during the second training session is not due to the specific training angle direction.

Finally, to determine whether the small directional bias plays any effects on learning the second task, we directly compared the initial deviation in the Perturb2 phase between *experiments 2* and *3* with an unpaired $t$-test. The result revealed no significant differences between them ($t_{21} = 0.5$, $P = 0.64$, $d = 0.20$), suggesting that there was no clear benefit of the small directional bias to the learning facilitation in the second training session.

Overall, *experiment 3* revealed results similar to *experiment 2*, that is, faster learning of the second task and its association with the increased exploratory behavior. These results indicate that faster relearning in the second training is not largely attributed to a potential directional bias and retrieval of the original memory.

### *Increased Exploration Continued Even After Longer Training of the Same Task*

In the prior experiments, we found that participants' motor exploration increased and remained elevated throughout the first task training. Thus we asked whether the magnitude changes with longer repetition of trials.

As found in the previous three experiments, participants' PA gradually shifted toward the predetermined target angle and converged on $17.2 \pm 1.5°$ in the last 40-trial epoch during the Perturb1 phase (Fig. 5*A*). This was accompanied by a gradual increase in the percentage of successful trials across epochs (Fig. 5*B*). During the subsequent Perturb2 phase, the shifted PA slightly improved and remained at the target until the end of the phase (last 40-trial epoch $19.4 \pm 0.7°$). As found in the prior experiments, the magnitude of |ΔPA| after failed trials showed gradual increases (Fig. 5*C*). Interestingly, this change remained high during the Perturb2 phase (first and last 40 trial epochs $6.4 \pm 1.1°$ and $6.4 \pm 0.8°$), despite the continuous trial repetition in the same task. These results indicate that the magnitude of motor exploration persisted elevated even during longer training of the same task. Of note, the percentage of task success trials did not show a clear return to baseline levels ($78.3 \pm 4.5\%$), even at the end of the Perturb2 phase (last epoch $64.8 \pm 4.9\%$, paired $t$-test, $t_9 = 2.2$, $P = 0.055$, $d_z = 0.70$).

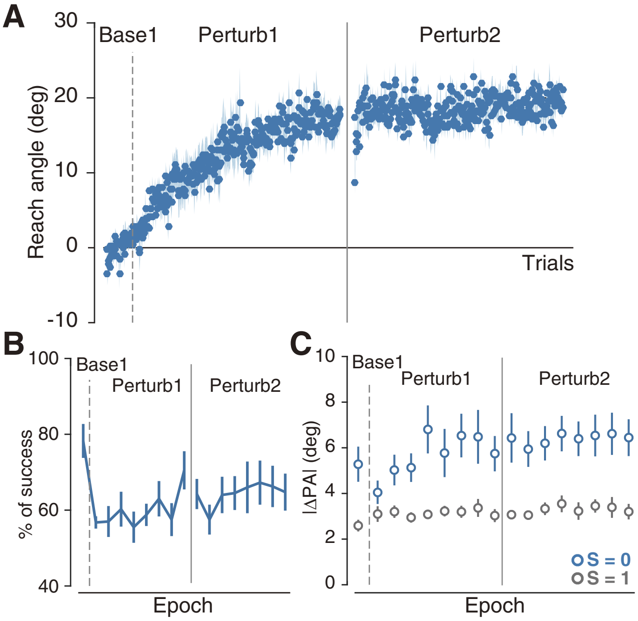

Fig. 5. Increased exploration remained elevated during longer repetition of trials. *A*: pointing angle (PA) while training the task. *B*: mean % of successful trials. *C*: unsigned trial-to-trial change in PA (|ΔPA|) after failure ($S = 0$, blue) and success ($S = 1$, gray) trials across epochs. Values show the mean and SE for each trial or 40-trial epoch.

### *Nonlearners Showed Little Exploratory Behavior Throughout Training*

We analyzed the Nonlearner group separately. This group constitutes the ideal control group to investigate whether the upregulated motor exploration during task training is associated with learning rather than simple task execution.

Task performance of the Nonlearner group during the Base1 phase was similar to that of the learners from the three experiments. Indeed, their PA converged on the visible target (PA 1.0 ± 0.9°; Fig. 6*A*) and resulted in high task accuracy (% of trial success 70.3 ± 4.8%; Fig. 6*B*). However, their PA did not show particular shifts toward the target angle when exposed to the Perturb1 phase, ending up with 5.8 ± 1.3° in the last 40-trial epoch (Fig. 6*A*). This was accompanied by no increase, but rather a decrease, in task accuracy throughout training (Fig. 6*B*). Additionally, unlike the learners, we did not find a gradual increase in the magnitude of |ΔPA| after either failed or successful trials throughout training {RM ANOVA [Outcome (2) × Epoch (8)], Outcome × Epoch interaction $F_{1.1,\ 10.2} = 0.9$, $P = 0.39$, $\eta_p^2 = 0.10$, effect for the factor Epoch $F_{1.2,\ 10.8} = 1.0$, $P = 0.36$, $\eta_p^2 = 0.10$; Fig. 6*C*}. This result supports the view that, rather than mere task execution, the learners' gradual increase in exploratory behavior during task training was largely attributed to learning. Furthermore, in the Nonlearner group, we found a magnitude of |ΔPA| after failed trials comparable to that after successful trials (effect for the factor Outcome $F_{1,\ 9} = 4.6$, $P = 0.06$, $\eta_p^2 = 0.34$). This result suggests that the Nonlearner group may have less sensitivity to the negative feedback than the learners.

To further investigate feedback sensitivity of the Nonlearner group and compare them to the learners, we extended trial-to-trial analysis for the magnitude of |ΔPA| to include the history of past feedback (Pekny et al. 2015). Here, we considered all eight possible combinations of success and failure feedback for three consecutive trials. The feedback history for three consecutive trials was represented by variables $S(n)$, $S(n - 1)$, and $S(n - 2)$, indicating whether task performance was successful in trials $n$, $n - 1$, and $n - 2$, respectively (Fig. 6*D*). For this analysis, we used trials only from the Base1 and Perturb1 phases that were completed by all the participants.

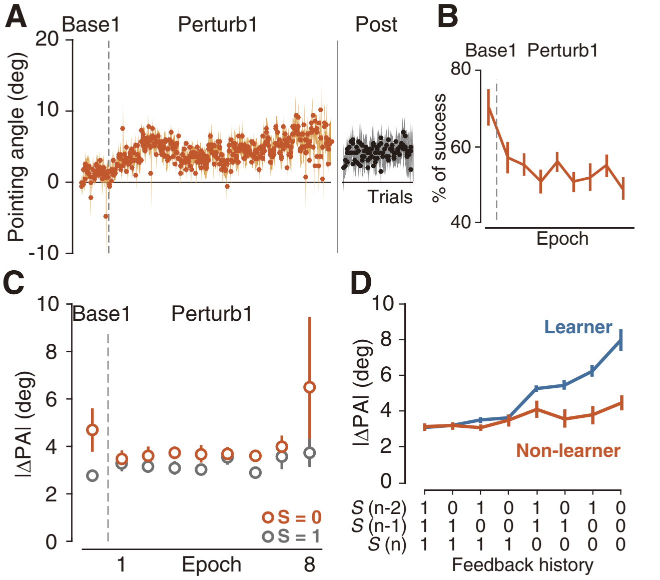

Fig. 6. Nonlearners showed no changes in exploratory behavior. *A*: pointing angle (PA) during task training. Dots and shaded area show the mean and SE for each trial during the Base1-Perturb1 (orange) and Post (black) phases. Only the data collected in common across all the participants (i.e., Base1, Perturb1, and Post) are presented. *B*: mean % of successful trials. C: unsigned trial-to-trial change in PA (|ΔPA|) for each epoch. Orange and gray dots indicate trials after failure ($S = 0$) and success ($S = 1$), respectively. *D*: |ΔPA| between trials $n$ and $n + 1$ as a function of the feedback history for the 3 most recent trials [$S$ ($n$), $S$ ($n - 1$), $S$ ($n - 2$)]. Orange and blue lines indicate data from the Nonlearner and Learner groups, respectively.

When we first analyzed |ΔPA| as a function of the feedback in the learners, we did not find significant differences among the participants across the four experiments {2-way mixed-effect RM ANOVA [History (8 combinations) × Experiment (*experiments 1*, *2*, *3*, and *4*)], effect for the factor Experiment ($F_{3,\ 44} = 1.3$, $P = 0.29$, $\eta_p^2 = 0.08$), History × Experiment interaction ($F_{4.5,\ 65.7} = 0.4$, $P = 0.83$, $\eta_p^2 = 0.03$}. Therefore, we grouped all the learners from the three experiments into one group (Learner group, $n = 38$) and compared them to the nonlearners. We found that |ΔPA| gradually increased as unsuccessful feedback history accumulated in the Learner group, whereas in the Nonlearner group it showed little increase in response to failure feedback accumulation {RM ANOVA [History (8) × Group (Learner, Nonlearner)], History × Group interaction $F_{1.6,\ 60.2} = 6.4$, $P = 0.005$, $\eta_p^2 = 0.10$; Fig. 6*D*}. This result indicates that the nonlearners have less sensitivity to unsuccessful feedback, which may lead to less exploratory behavior and learning.

## DISCUSSION

Exploring for the correct motor actions is a critical element of reinforcement motor learning. However, whether and how the exploratory behavior changes as individuals learn a new task has not been well characterized. As predicted, we found that participants' motor exploration gradually increased as they trained on a reinforcement-based motor task. However, the exploratory behavior remained elevated even after clear improvements, and stability, in task execution accuracy. Furthermore, participants showed sustained increased motor exploration when they were exposed to a second bout of training in the same task. This effect was associated and proportional to faster relearning. In contrast, participants who were unable to sufficiently learn the task demonstrated few changes in exploratory behavior. This finding confirmed that exploration is critical to reinforcement learning and indicates that the gradual increase in motor exploration is not the result of mere task execution. In addition, our findings suggest that the motor system can upregulate the amount of motor exploration during learning a reinforcement-based motor task as if acquiring a beneficial strategy that facilitates subsequent learning.

In the motor domain, reinforcement learning has been described as the process in which actions leading to successful outcomes are reinforced while those leading to unsuccessful outcomes are avoided (Sutton and Barto 1998). This form of learning necessitates exploratory behavior through which the motor system identifies or updates values of potential actions in a trial-by-trial manner based on each action outcome. Therefore, as observed here, an increase in motor exploration should be expected during reinforcement learning so that the motor system can identify the reward landscape in action space.

However, previously it was not clear whether the increased exploratory behavior returns to baseline levels once the task has been learned (consistent high level of performance success).

In this study, we found that the amount of exploration remained elevated throughout training. This was observed even when the percentage of successful trials continued to increase throughout training. At first glance, this result might seem contradictory to previous findings indicating that the amount of exploratory behavior can be regulated by an overall rewarding situation, e.g., the probability of reward. For instance, studies across a variety of species indicated that the overall trial-to-trial change in motor output tends to decrease under a condition where the probability for positive reward outcome increases (Galea et al. 2013; Gharib et al. 2001, 2004; Pekny et al. 2015; Stahlman et al. 2010; Stahlman and Blaisdell 2011; Takikawa et al. 2002). There are some important differences between those studies and ours. First, we measured trial-to-trial movement changes after failed trials as a proxy for exploratory behavior, whereas previous investigations measured trial-to-trial changes regardless of the previous outcome (i.e., overall variability). Second, we tracked changes in exploration as a function of learning-associated changes in reward, whereas previously overall variability was determined in separate blocks across a variety of experimentally controlled rewarding conditions. Nevertheless, these previous findings showing the overall reduction in movement-to-movement variability do not contradict that changes after failure increase in magnitude and remain elevated in the presence of errors.

We found that the exploratory behavior started at a greater magnitude from the onset of the second training compared with the first training, even when the movement directions were different. This initial increase in exploration was associated with, and proportional to, faster learning of the second task relative to the first task. These results can be interpreted as the motor system not only learning a new task-specific motor pattern but also gaining the strategy of being more exploratory. In other words, it acquires knowledge on how to learn different tasks given the same context (e.g., learning to learn or meta-learning; Braun et al. 2009, 2010; Krakauer and Mazzoni 2011) to increase the efficiency of subsequent training. However, we cannot conclude whether the motor system actually "learned" the strategy or remained in a heightened exploratory state since the percentage of task success remained low. To determine this, in *experiment 4* participants repeated more training trials, to evaluate time course changes in the magnitude of exploration while the success rate becomes substantially higher. We found, however, that the rate of task success still remained lower than baseline even after longer repetition of trials. Thus we were unable to dissociate the effects leading to a heightened exploration; a lower success rate may have kept motor exploration greater, or alternatively greater exploration (i.e., greater magnitude of movement changes after failure) as a learned strategy may have resulted in a certain probability of task failure (i.e., lower success rate). Future studies should consider experimental designs to disentangle this relationship.

The present study cannot clarify whether the increased exploratory behavior represents "explicit" strategic changes that participants intentionally controlled. Interestingly, our results showed different patterns of motor performance during the Post phase (i.e., no-feedback trials) among the three experiments, such as a distinct drop relative to the end of the Perturb1 phase and/or an upward drift (Figs. 2*A*, 3*A*, and 4*A*). Although these results were somewhat unexpected because all the participants experienced exactly the same training schedule until the end of the Post phase, these seemingly contradictory results may be due to a different degree of dependence on explicit components across individuals during task training or during the subsequent Post phase. Indeed, a recent human behavioral study suggests that explicit strategy is partly engaged in the process of reinforcement learning, since motor patterns learned through reinforcement mechanisms were degraded when the use of explicit strategy was experimentally constrained or intentionally removed (Holland et al. 2018). On the other hand, generating identical movements in successive trials is virtually impossible. In other words, movements exhibit trial-to-trial variability regardless of outcomes in a preceding trial. This type of movement variability or "motor noise" is thought to be, in part, an inherent feature originated from the fundamental properties of the neuromuscular system (Dhawale et al. 2017; Faisal et al. 2008; Jones et al. 2002; van Beers et al. 2004). Therefore, movement changes in response to failed outcomes include motor noise-related changes as well as exploratory action changes. We think that the contribution of motor noise to the gradual increase in motor exploration after failed trials during training, however, is very limited. This is because motor noise was also present after successful trials and in the Nonlearner group, yet these two situations did not lead to changes in exploratory behavior. Thus we interpret the increase in exploration after failed trials during learning as largely related to an exploratory strategy, irrespective of the level of awareness, rather than simple motor noise.

The increased exploratory behavior led to faster learning in subsequent practice exposures, a phenomenon known as savings (Krakauer et al. 2005; Mawase et al. 2014, 2017a; Zarahn et al. 2008). Although debate continues as to which mechanisms contribute to savings (Huang et al. 2011; Leow et al. 2016; Morehead et al. 2015; Orban de Xivry and Lefèvre 2015; Roemmich and Bastian 2015), a study suggested the potential contribution of movement direction bias resulting from successful movement repetitions (Huang et al. 2011). In addition, it is possible that the observed small biases in movement correction direction may have also contributed to faster learning, although this hypothesis cannot explain why participants experienced faster relearning in *experiment 3*, where the biases were in the opposite direction to the appropriate movement corrections. The faster relearning in *experiment 3* also cannot be attributed to prior exposure to the Wash phase, where participants experienced movements in the same rotation direction as in the second training. This is because any tangible effect should then have been present in *experiment 2*, where participants experienced a washout in the opposite direction to the second training yet no negative effect of bias was observed in the second training exposure. Therefore, direction biases cannot fully explain the faster relearning observed in the second training.

Our *experiment 3* results further posit that the effect of increased motor exploration can be generalized to another task that requires a different motor solution but relies on the same reinforcement learning paradigm. Therefore, it is conceivable that training on reinforcement-based tasks could be used as a primer to increase exploratory behavior and enhance the efficiency of subsequent learning in the same contextual setting. It could be argued that our findings are specific to the nature of our present experimental paradigm testing reinforcement learning. Here we used a task in which the target direction cannot be learned in any other ways except through exploration and reinforcing feedback. Therefore, the present findings might only be applicable to reinforcement forms of learning. Generalization to other learning forms (e.g., error-based learning, use-dependent learning) should be done with caution and formally tested in future studies.

Interestingly, some participants were unable to reach the optimal motor pattern during the first training session. These individuals did not show a gradual increase in exploratory behavior throughout training. This group of participants seems to have less sensitivity to binary feedback, particularly to negative feedback. Although it cannot be distinguished whether the reduced susceptibility is accounted for by the level of perception or the level of motor planning integration, the finding can explain why the nonlearners exhibited less learning efficiency and therefore no gradual increase in exploration during training.

In summary, the present study demonstrates that the motor system upregulates the amount of motor exploration during reinforcement learning. In turn, this effect is associated with improvements in learning efficiency during subsequent training in the same contextual setting. Although we cannot be conclusive, our findings might suggest that the motor system acquires knowledge to become more exploratory as if developing an overall strategy that can be useful to learn new motor actions, at least in the presence of similar context. The present results open an opportunity to design better training paradigms for healthy individuals as well as for people undergoing rehabilitation to speed up learning.

## ACKNOWLEDGMENTS

We thank Tziporah Thompson for the illustration in Fig. 1*A*.

Present address for S. Uehara: Faculty of Rehabilitation, Fujita Health University School of Health Sciences.

Present address for F. Mawase: Faculty of Biomedical Engineering, Technion—Israel Institute of Technology, Israel.

## GRANTS

This work was supported by National Institutes of Health Grant 4R01 HD-073147-05 (to P. Celnik) and a Japan Society for the Promotion of Science Overseas Research Fellowship (to S. Uehara).

## DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

## AUTHOR CONTRIBUTIONS

S.U., F.M., A.S.T., and P.A.C. conceived and designed research; S.U., F.M., and K.M.C.-A. performed experiments; S.U. and F.M. analyzed data; S.U., and P.A.C. interpreted results of experiments; S.U. prepared figures; S.U. drafted manuscript; S.U., F.M., A.S.T., K.M.C.-A., and P.A.C. edited and revised manuscript; S.U., F.M., A.S.T., K.M.C.-A., and P.A.C. approved final version of manuscript.

## REFERENCES

**Aronov D, Andalman AS, Fee MS.** A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* 320: 630–634, 2008. doi:10.1126/science.1155140.

**Braun DA, Aertsen A, Wolpert DM, Mehring C.** Motor task variation induces structural learning. *Curr Biol* 19: 352–357, 2009. doi:10.1016/j.cub.2009.01.036.

**Braun DA, Mehring C, Wolpert DM.** Structure learning in action. *Behav Brain Res* 206: 157–165, 2010. doi:10.1016/j.bbr.2009.08.031.

**Chen X, Mohr K, Galea JM.** Predicting explorative motor learning using decision-making and motor noise. *PLoS Comput Biol* 13: e1005503, 2017. doi:10.1371/journal.pcbi.1005503.

**Dhawale AK, Smith MA, Ölveczky BP.** The role of variability in motor learning. *Annu Rev Neurosci* 40: 479–498, 2017. doi:10.1146/annurev-neuro-072116-031548.

**Diedrichsen J, White O, Newman D, Lally N.** Use-dependent and error-based learning of motor behaviors. *J Neurosci* 30: 5159–5166, 2010. doi:10.1523/JNEUROSCI.5406-09.2010.

**Faisal AA, Selen LP, Wolpert DM.** Noise in the nervous system. *Nat Rev Neurosci* 9: 292–303, 2008. doi:10.1038/nrn2258.

**Fiete IR, Fee MS, Seung HS.** Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *J Neurophysiol* 98: 2038–2057, 2007. doi:10.1152/jn.01311.2006.

**Galea JM, Ruge D, Buijink A, Bestmann S, Rothwell JC.** Punishment-induced behavioral and neurophysiological variability reveals dopamine-dependent selection of kinematic movement parameters. *J Neurosci* 33: 3981–3988, 2013. doi:10.1523/JNEUROSCI.1294-12.2013.

**Garst-Orozco J, Babadi B, Ölveczky BP.** A neural circuit mechanism for regulating vocal variability during song learning in zebra finches. *eLife* 3: e03697, 2014. doi:10.7554/eLife.03697.

**Gharib A, Derby S, Roberts S.** Timing and the control of variation. *J Exp Psychol Anim Behav Process* 27: 165–178, 2001. doi:10.1037/0097-7403.27.2.165.

**Gharib A, Gade C, Roberts S.** Control of variation by reward probability. *J Exp Psychol Anim Behav Process* 30: 271–282, 2004. doi:10.1037/0097-7403.30.4.271.

**Holland P, Codol O, Galea JM.** Contribution of explicit processes to reinforcement-based motor learning. *J Neurophysiol* 119: 2241–2255, 2018. doi:10.1152/jn.00901.2017.

**Huang VS, Haith A, Mazzoni P, Krakauer JW.** Rethinking motor learning and savings in adaptation paradigms: model-free memory for successful actions combines with internal models. *Neuron* 70: 787–801, 2011. doi:10.1016/j.neuron.2011.04.012.

**Izawa J, Shadmehr R.** Learning from sensory and reward prediction errors during motor adaptation. *PLoS Comput Biol* 7: e1002012, 2011. doi:10.1371/journal.pcbi.1002012.

**Jones KE, Hamilton AF, Wolpert DM.** Sources of signal-dependent noise during isometric force production. *J Neurophysiol* 88: 1533–1544, 2002. doi:10.1152/jn.2002.88.3.1533.

**Krakauer JW, Ghez C, Ghilardi MF.** Adaptation to visuomotor transformations: consolidation, interference, and forgetting. *J Neurosci* 25: 473–478, 2005. doi:10.1523/JNEUROSCI.4218-04.2005.

**Krakauer JW, Mazzoni P.** Human sensorimotor learning: adaptation, skill, and beyond. *Curr Opin Neurobiol* 21: 636–644, 2011. doi:10.1016/j.conb.2011.06.012.

**Leow LA, de Rugy A, Marinovic W, Riek S, Carroll TJ.** Savings for visuomotor adaptation require prior history of error, not prior repetition of successful actions. *J Neurophysiol* 116: 1603–1614, 2016. doi:10.1152/jn.01055.2015.

**Mawase F, Bar-Haim S, Shmuelof L.** Formation of long-term locomotor memories is associated with functional connectivity changes in the cerebellar-thalamic-cortical network. *J Neurosci* 37: 349–361, 2017a. doi:10.1523/JNEUROSCI.2733-16.2016.

**Mawase F, Shmuelof L, Bar-Haim S, Karniel A.** Savings in locomotor adaptation explained by changes in learning parameters following initial adaptation. *J Neurophysiol* 111: 1444–1454, 2014. doi:10.1152/jn.00734.2013.

**Mawase F, Uehara S, Bastian AJ, Celnik P.** Motor learning enhances use-dependent plasticity. *J Neurosci* 37: 2673–2685, 2017b. doi:10.1523/JNEUROSCI.3303-16.2017.

**Morehead JR, Qasim SE, Crossley MJ, Ivry R.** Savings upon re-aiming in visuomotor adaptation. *J Neurosci* 35: 14386–14396, 2015. doi:10.1523/JNEUROSCI.1046-15.2015.

**Orban de Xivry JJ, Lefèvre P.** Formation of model-free motor memories during motor adaptation depends on perturbation schedule. *J Neurophysiol* 113: 2733–2741, 2015. doi:10.1152/jn.00673.2014.

**Pekny SE, Izawa J, Shadmehr R.** Reward-dependent modulation of movement variability. *J Neurosci* 35: 4015–4024, 2015. doi:10.1523/JNEUROSCI.3244-14.2015.

**Roemmich RT, Bastian AJ.** Two ways to save a newly learned motor pattern. *J Neurophysiol* 113: 3519–3530, 2015. doi:10.1152/jn.00965.2014.

**Sidarta A, Vahdat S, Bernardi NF, Ostry DJ.** Somatic and reinforcement-based plasticity in the initial stages of human motor learning. *J Neurosci* 36: 11682–11692, 2016. doi:10.1523/JNEUROSCI.1767-16.2016.

**Stahlman WD, Blaisdell AP.** The modulation of operant variation by the probability, magnitude, and delay of reinforcement. *Learn Motiv* 42: 221–236, 2011. doi:10.1016/j.lmot.2011.05.001.

**Stahlman WD, Roberts S, Blaisdell AP.** Effect of reward probability on spatial and temporal variation. *J Exp Psychol Anim Behav Process* 36: 77–91, 2010. doi:10.1037/a0015971.

**Sutton RG, Barto AG.** *An Introduction to Reinforcement Learning*. Cambridge, MA: MIT Press, 1998.

**Takikawa Y, Kawagoe R, Itoh H, Nakahara H, Hikosaka O.** Modulation of saccadic eye movements by predicted reward outcome. *Exp Brain Res* 142: 284–291, 2002. doi:10.1007/s00221-001-0928-1.

**Therrien AS, Wolpert DM, Bastian AJ.** Effective reinforcement learning following cerebellar damage requires a balance between exploration and motor noise. *Brain* 139: 101–114, 2016. doi:10.1093/brain/awv329.

**Tumer EC, Brainard MS.** Performance variability enables adaptive plasticity of "crystallized" adult birdsong. *Nature* 450: 1240–1244, 2007. doi:10.1038/nature06390.

**Uehara S, Mawase F, Celnik P.** Learning similar actions by reinforcement or sensory-prediction errors rely on distinct physiological mechanisms. *Cereb Cortex* 28: 3478–3490, 2018. doi:10.1093/cercor/bhx214.

**van Beers RJ, Haggard P, Wolpert DM.** The role of execution noise in movement variability. *J Neurophysiol* 91: 1050–1063, 2004. doi:10.1152/jn.00652.2003.

**Verstynen T, Sabes PN.** How each movement changes the next: an experimental and theoretical study of fast adaptive priors in reaching. *J Neurosci* 31: 10050–10059, 2011. doi:10.1523/JNEUROSCI.6525-10.2011.

**Wu HG, Miyamoto YR, Gonzalez Castro LN, Ölveczky BP, Smith MA.** Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nat Neurosci* 17: 312–321, 2014. doi:10.1038/nn.3616.

**Zarahn E, Weston GD, Liang J, Mazzoni P, Krakauer JW.** Explaining savings for visuomotor adaptation: linear time-invariant state-space models are not sufficient. *J Neurophysiol* 100: 2537–2548, 2008. doi:10.1152/jn.90529.2008.