Reward feedback accelerates motor learning

Ali A. Nikooyan and Alaa A. Ahmed

J Neurophysiol 113:633-646, 2015. First published 29 October 2014; doi:10.1152/jn.00032.2014

J Neurophysiol 113: 633–646, 2015.
First published October 29, 2014; doi:10.1152/jn.00032.2014.

Reward feedback accelerates motor learning

Ali A. Nikooyan and Alaa A. Ahmed
Department of Integrative Physiology, University of Colorado Boulder, Boulder, Colorado

Submitted 10 January 2014; accepted in final form 26 October 2014

Nikooyan AA, Ahmed AA. Reward feedback accelerates motor learning. *J Neurophysiol* 113: 633–646, 2015. First published October 29, 2014; doi:10.1152/jn.00032.2014.—Recent findings have demonstrated that reward feedback alone can drive motor learning. However, it is not yet clear whether reward feedback alone can lead to learning when a perturbation is introduced abruptly, or how a reward gradient can modulate learning. In this study, we provide reward feedback that decays continuously with increasing error. We asked whether it is possible to learn an abrupt visuomotor rotation by reward alone, and if the learning process could be modulated by combining reward and sensory feedback and/or by using different reward landscapes. We designed a novel visuomotor learning protocol during which subjects experienced an abruptly introduced rotational perturbation. Subjects received either visual feedback or reward feedback, or a combination of the two. Two different reward landscapes, where the reward decayed either linearly or cubically with distance from the target, were tested. Results demonstrate that it is possible to learn from reward feedback alone and that the combination of reward and sensory feedback accelerates learning. An analysis of the underlying mechanisms reveals that although reward feedback alone does not allow for sensorimotor remapping, it can nonetheless lead to broad generalization, highlighting a dissociation between remapping and generalization. Also, the combination of reward and sensory feedback accelerates learning without compromising sensorimotor remapping. These findings suggest that the use of reward feedback is a promising approach to either supplement or substitute sensory feedback in the development of improved neurorehabilitation techniques. More generally, they point to an important role played by reward in the motor learning process.

reinforcement learning; dopaminergic; decision-making; sensorimotor mapping; temporal-difference model

REWARD-BASED LEARNING has been at the forefront of advances in many disciplines, ranging from psychology (Dydewalle 1982) to artificial intelligence and machine learning (Kaelbling et al. 1996), robotics (Kormushev et al. 2013), and, most recently, neuroeconomics (Glimcher et al. 2009). In contrast to its fast-growing trend in these disciplines, reward-based learning has received less attention in the study of how the brain learns new movements. In a sequential key-pressing task, it was shown (Palminteri et al. 2011) that positive (monetary) reward could improve learning in patients with Tourette syndrome (which is thought to be related to hyperactivity of the dopaminergic transmission). Another study also demonstrated that monetary reward leads to improvements in motor memory in healthy adults during performance of a skill-learning isometric force task (Abe et al. 2011).

Although these studies support a role for reward in motor learning, its role in tasks involving motor adaptation is less clear. Motor adaptation is a form of motor learning that is driven by sensory prediction error and leads to an update of the sensorimotor mapping, or forward model, between the limbs and the environment. Adaptation tasks involve gradual improvement in performance over time in response to a change in the environment (Krakauer and Mazzoni 2011). Popular paradigms to investigate this process impose perturbations on subjects' reaching movements via either a visuomotor rotation (Krakauer et al. 2000; Nikooyan and Zadpoor 2009) or a force field (Ahmed and Wolpert 2009; Huang and Ahmed 2013; Huang et al. 2012). These studies, however, have mostly focused on the process of learning from sensory feedback by quantifying the resultant changes in the sensory mapping between the limb and the external environment.

Izawa and Shadmehr (2011) tried to distinguish between learning from sensory and reward feedback during a visuomotor rotation arm-reaching task. Despite showing greater endpoint variability, people were able to learn from reward feedback alone and to a level comparable to that learned from sensory feedback. Their results also revealed that reward-based learning was fundamentally distinct from sensory feedback-based learning. Namely, learning from reward alone did not lead to an update of a sensorimotor map of the relationship between the arm and cursor position or generalization of learning to nearby target directions, as is normally observed when learning a visuomotor rotation. The authors proposed a two-component additive learning process: action-selection and the learning of a sensorimotor map, or internal model of the relationship between arm and cursor position. Within this framework, learning from reward feedback alone does not allow for internal model learning, so the amount of internal model learning will be directly related to the quality of the sensory feedback and inversely related to the degree of action selection.

However, there were certain details of the study that make it difficult to extend the results and thus leave many open questions regarding reward-based learning. First, the visuomotor rotation was introduced gradually, and with limited size (up to 8°), rather than the more canonical abrupt rotation of ~30° (Krakauer et al. 2000). Second, in their study only binary reward was provided depending on whether the trial was successful or not.

Recent studies have revealed that changes in the central nervous system could depend on the manner of introducing the perturbations. Schlerf et al. (2012) found that the level of cerebellar inhibition would increase when the perturbation was abruptly introduced during a visuomotor rotation task, whereas little change was observed with a gradual perturbation. They came to the conclusion that neural bases of learning from abrupt perturbations and from gradual perturbations are distinct. These findings imply that reward-based learning may

Address for reprint requests and other correspondence: A. A. Ahmed, Neuromechanics Lab, Clare Small 106, 1725 Pleasant St., UCB 354, Dept. of Integrative Physiology, Univ. of Colorado Boulder, Boulder, CO 80309-0354 (e-mail: alaa@colorado.edu).

also differ in response to a gradual compared with an abrupt perturbation.

From an ecological standpoint, many of the errors experienced in daily life are typically abrupt, not gradual. We inevitably experience large errors, not just small ones. Furthermore, reward is mostly a continuous signal that can have a range of values and may not be limited to binary ones. Using binary reward could also limit the experiment to only small perturbations because of the difficulty of trying to adapt to a larger perturbation with only yes/no feedback. There are studies that have used continuous reward feedback, rather than binary feedback, in motor learning (Dam et al. 2013; Hoffman et al. 2008). In those studies, the aim was to match a hidden target/target trajectory. Reward was a continuous signal that provided subjects with information about how closely they could reach a hidden target (trajectory). However, these studies did not investigate the underlying representation of learning, and it is not clear how their findings will generalize to the type and magnitude of error experienced in stereotypical visuomotor rotation tasks.

Taking these findings together, it has yet to be determined whether it is possible to learn in response to large errors using reward feedback alone. This is the first question (Q1) we seek to answer with this study. In addition to its potential role as a substitute for sensory feedback, there is some evidence that reward as a supplementary form of feedback can improve learning in terms of reducing learning variability (Izawa and Shadmehr 2011; Manley et al. 2014). As such, our second question (Q2) is whether the learning process can be modulated by combining both reward and sensory feedback. Moreover, studies in the field of artificial intelligence have shown that using alternate reward landscapes could, theoretically, accelerate the learning process (Mataric 1994; Niekum et al. 2011). Thus we also seek to investigate to what extent the learning process in human subjects can be modulated with different reward landscapes. Finally, we hope to understand how reward feedback influences the relative contributions of sensorimotor remapping and action selection to the overall learning process, thereby shedding light on the underlying neural mechanisms.

To pursue these questions, a novel experimental protocol was designed during which subjects experienced an abruptly introduced visuomotor rotation of significant size and received visual feedback alone, reward feedback alone, or a combination of both visual and reward feedback. Instead of a binary reward, continuous reward feedback (i.e., a reward gradient) was presented to the subjects in the form of trial score. We tested subjects in a linear reward landscape, where the reward decayed linearly with distance from the target, and in a cubic landscape, where the reward decayed more steeply with distance from the target. In a second set of experiments, we investigated the effects of reward feedback on internal model learning and action selection by quantifying the degree of sensorimotor remapping and generalization of learning to nearby targets.

MATERIALS AND METHODS

Statement of Ethics

The Institutional Review Board of the University of Colorado Boulder approved the experimental procedure. All subjects agreed to participate by providing informed consent. Subjects reported no history of neurological or neuromuscular diseases.

Experimental Protocol

Experiment 1: setup. Subjects ($n = 46$, recruited through the University of Colorado Boulder Psychology 1001 Subject Pool) were seated in a chair with full back support and in front of a robotic manipulandum (Interactive Motion Technologies shoulder-elbow robot 2; Fig. 1*A*). Trunk movement was limited by use of shoulder straps and a lap belt (Fig. 1*A*). A flat-screen liquid crystal display (LCD) monitor was mounted in front of the subjects at eye-level (Fig. 1*A*). Subjects were all right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield 1971). While grasping the handle of the robotic arm with their right hand, subjects were instructed to make 15-cm rapid out-and-back horizontal reaching movements to move an on-screen cursor ($r = 0.3$ cm) from a home circle ($r = 2$ cm) near the bottom of the monitor to a rectangular target placed on a target arc ($r = 15$ cm, thickness = 0.3 cm, color: green) near the top of the monitor and then return to the home circle (Fig. 1*B*). Every trial started by repositioning the cursor within the home circle and was completed when the cursor reached the target arc (Fig. 1*B*, *left*). Subjects were instructed to make a rapid movement to the target, i.e., neither "too fast" nor "too slow." If the outward movement took place within the required time limit (200–600 ms), the target arc "exploded"; otherwise, it turned gray if the movement was too slow or red if the movement was too fast. Importantly, the subjects were not required to settle in the target arc, but merely to "hit" it. During each trial, the position of the cursor's *x-y* coordinates (Fig. 1*B*) was sampled at 200 Hz. Visual feedback of the subjects' hand was occluded with a horizontal, opaque screen (Fig. 1*B*, *left*).

We designed an experiment that considered three reward landscapes (none, linear, or cubic) and two sensory feedback states (visual feedback or no visual feedback), producing a 3×2 factorial design, as illustrated in Fig. 1*C*. Our primary goal was to quantify the effects of reward and sensory feedback on this learning process.

The experiment consisted of 600 trials (Fig. 1*D*) beginning with 50 familiarization trials in which all subjects received visual feedback of the cursor position. Familiarization trials were followed by a baseline block consisting of 50 trials during which subjects either received visual feedback (Vision groups) or no visual feedback (No-Vision groups) as per their group assignment (Fig. 1*C*). As the subjects made the out-and-back movement, the motion of the cursor underwent a 30° counterclockwise abrupt visuomotor rotational perturbation with respect to the motion of the hand (−30°), requiring the subjects to learn to compensate for this perturbation (rotation block). We used this paradigm to examine how reward and sensory feedback interacted to modulate learning. After 450 rotation trials, the environment returned to a 0° rotation for the remaining 50 trials (washout block). A 30-s rest period was provided every 200 trials (but not between different blocks; Fig. 1*D*). Subjects were told that at some point during the experiment, the task might become more difficult; no additional information about the timing and/or the type of difficulty was provided.

During the baseline, rotation, and washout portions of the experiment, the subjects in the No-Vision groups could only rely on reward feedback to learn the rotation (they could only see the home circle at movement onset in each trial), whereas subjects in the Vision groups received visual feedback. In the groups where reward feedback was provided, it was presented as soon as the subjects reached to any point on the target arc (regardless of the cursor being visible to them or not) and was reported as a trial score that ranged from 0 to 1,000 (Fig. 1*E*). Depending on the reward group (Linear, Cubic), the trial score (R) depended on the following function:

$$R = 1,000 \times \left(\frac{180 - |\theta|}{180} \right)^{\alpha}, \quad (1)$$

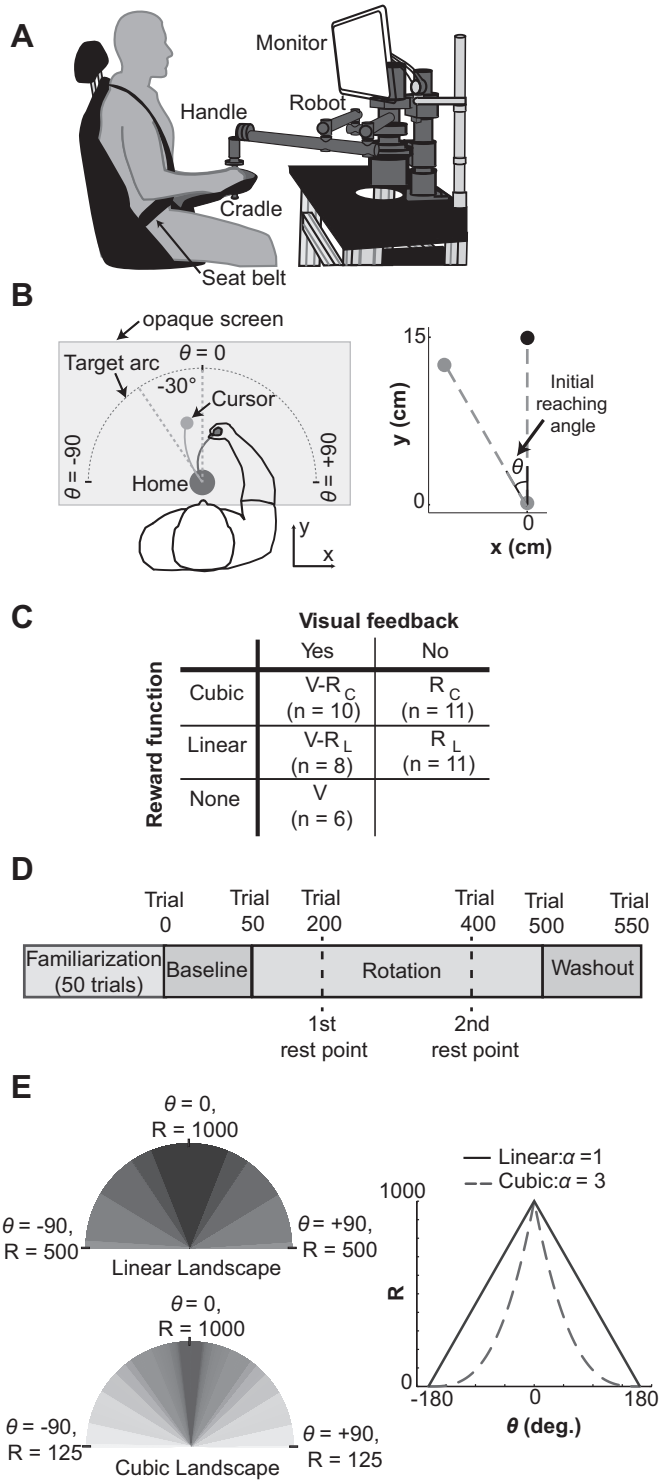

Fig. 1. *Experiment 1*: description. *A*: robotic arm and monitor (opaque screen not shown). *B*: reaching movement from the home circle to the target arc (*left*) and typical reaching trajectory (in the *x*-*y* plane) in the presence of a counterclockwise rotational perturbation (*right*). *C*: design of experimental groups for *experiment 1*. Vision groups received visual feedback with no reward (V; Control group) or with cubic (V-R_C) or linear (V-R_L) reward; No-Vision groups had no visual feedback but received cubic (R_C) or linear (R_L) reward. *D*: timeline of experimental blocks for *experiment 1*. *E*: visual representation of the linear and cubic reward landscapes. R, trial score; θ, reaching angle.

where θ is the initial cursor reaching angle in degrees (Fig. 1*B*, *right*) and was calculated at the time where the Euclidian distance between the centers of the cursor and the home circle first exceeded 3 cm (~100 ms from movement onset).

Two different reward landscapes were used: linear and cubic. In the linear landscape, the trial score decayed linearly ($\alpha = 1$) with the initial reaching angle (Fig. 1*E*). In the cubic landscape, the trial score decayed cubically ($\alpha = 3$) with the initial reaching angle (Fig. 1*E*). In other words, the cubic landscape penalized a given angular error much more strongly than the linear landscape. The rationale behind choosing these reward landscapes was, first, they provided a simple and intuitive way to penalize errors differentially, and second, preliminary modeling results suggested that these two landscapes would lead to different and distinguishable learning performance. The value of the obtained trial score together with its maximum possible value (equal to 1,000) was shown to the subjects on completion of each trial. The calculated value of the trial score (R; *Eq. 1*) was rounded to the closest integer value before being shown to the subjects. For instance, with a 30° initial reaching angle (i.e., $\theta = 30°$), the trial score for the linear landscape was shown as "833 out of 1,000," whereas that for the cubic landscape was given as "579 out of 1,000." Subjects receiving reward feedback were instructed to maximize their trial score.

Experiment 1: groups. In total, five groups were tested to investigate the effect of reward and visual feedback on learning (Fig. 1*C*). Each performed the experiment under a unique combination of visual feedback (Vision, No-Vision) and reward feedback (none, Linear, Cubic). The Vision groups are the V, V-R_C, and V-R_L groups, i.e., all the groups that received visual feedback. The V group served as a Control group that received visual feedback and no reward feedback. Hence, this group experienced an environment analogous to that experiment in standard visuomotor adaptation tasks. The V-R_C group received visual feedback and cubic reward feedback. The V-R_L group received visual feedback and linear reward feedback. Sometimes we refer to the V-R_L and the V-R_C groups together and call them the "Vision-Reward" groups. The No-Vision groups are the R_C and R_L groups, i.e., the groups that did not receive any visual feedback. The R_C group received cubic reward feedback and no visual feedback. The R_L group received linear reward feedback and no visual feedback. The groups can also be classified by the reward landscape experienced. R_C and V-R_C are the Cubic groups. R_L and V-R_L are the Linear groups.

Experiment 1: data analysis. To quantify the effects of reward and sensory feedback on the learning process, we compared learning between the five groups on the basis of three metrics: error, learning rate, and variability.

ERROR. Error on a given trial was taken to be the initial cursor reaching angle (θ; Fig. 1*B*). Error was averaged over bins of five trials and then averaged over all subjects in that group. Based on the initial reaching angle, the trial score (R) was also calculated from *Eq. 1* and averaged over each bin. The first 5 trials (i.e., the first bin) and the last 10 trials (i.e., the last 2 bins) at each experimental block (including baseline, rotation, and washout) were defined as the early and the late phases of that block, respectively.

To assess learning in different experimental groups, mean error at the early and the late baseline, learning (rotation), and washout phases was compared across all groups. To this end, a three-way ANOVA ($3 \times 2 \times 6$) was carried out where the main effects and interactions of reward landscape, visual feedback, and phase were evaluated. Here, reward landscape and visual feedback are between-subjects factors and phase is a within-subjects factor. A separate analysis was also carried out to compare mean trial score at the early and late learning phases for subjects in the No-Vision and the Vision-Reward groups.

LEARNING RATE. Previous studies have shown that error in the early phase of learning a visuomotor rotation can be characterized well with a single-rate exponential function (Zarahn et al. 2008). Thus the rate constant (c) of an exponential function (f) fit to the error data

was used as a measure for between-group comparisons of learning rate:

$$f(x) = a + be^{-cx}, \quad (2)$$

Based on the observed trend in our experimental data, we fit the first 50 trials of the learning block, similar to the first 30 trials used by Zarahn et al. (2008). Unconstrained nonlinear optimization algorithm was applied to find the exponential fit to the individual subject data as well as the mean error data for each group. A custom MATLAB (version R2013a; The MathWorks, Natick, MA) code was developed in which the "fminsearch" algorithm was applied for optimization. For the individual subject fits, parameters were compared using independent *t*-tests. For the fits to the average data, the statistics were computed differently: first, 95% confidence intervals for the nonlinear fit parameters were calculated using the "nlparci" MATLAB algorithm. Separately for each coefficient (including *a*, *b*, and *c* in *Eq. 2*), the confidence intervals were then compared between the fits to average data in each group. Nonoverlapping confidence intervals would necessarily indicate a significant difference, whereas for overlapping confidence intervals we used the following equation (Wolfe and Hanley 2002) to determine significance:

$$\left|\text{mean}_A - \text{mean}_B\right| > 2\sqrt{\left(\text{SE}_A^2 + \text{SE}_B^2\right)}, \quad (3)$$

where SE is the standard error of the mean. The difference between the two means with overlapping confidence intervals is significant only if the above inequality holds true.

When an exponential fit was not appropriate, we used an alternative method to evaluate the learning rate. For those groups, we compared the learning rate on the basis of the mean error value (at the interval of interest), where the smaller error would indicate a faster learning rate. Three different intervals within the rotation block were considered for comparison: the 1st interval comprised the 1st 50 trials, the 2nd interval was from *trial 51* to *200*, and the 3rd interval was from *trial 201* to *450*. Separately at each defined interval, mean errors across all subjects in each group were compared with each other.

To compare learning from reward feedback alone to learning with visual feedback alone, the No-Vision groups were compared with the Control group, V. To quantify the effect of different reward landscapes when no visual feedback was provided, the No-Vision groups were compared with each other. To quantify the effect of pairing reward feedback with visual feedback on learning rate, we calculated the combined learning rate in the Vision-Reward (i.e., V-R_C and V-R_L) groups and compared it with that in the Control group (V). Additionally, we compared learning rates in the V-R_C and V-R_L groups to determine the effect of reward landscape when visual feedback was present.

VARIABILITY. Reach variability in the first 100 trials (early learning variability) and the last 100 trials (late learning variability) in the rotation block was also calculated for each subject and compared between groups. The standard deviation of the error across 100 trials was taken as a measure of variability at each stage. Error decreases more rapidly early in adaptation; this faster drop could inflate the standard deviation. Therefore, we detrended the data to remove the mean using the "detrend" function in MATLAB. Variability was then calculated as the standard deviation of the detrended error data at each stage. A three-way ANOVA ($3 \times 2 \times 2$) was carried out where the main effects and interactions of reward landscape, visual feedback, and phase were evaluated.

Experiment 2: setup. This experiment was designed to examine the underlying mechanisms of learning from different forms of feedback via the addition of localization and generalization probe trials. The experimental apparatus and protocol were the same as in *experiment 1* (Fig. 1*A*). *Experiment 2* consisted of 875 trials (Fig. 2*A*) beginning with 70 familiarization trials during which all subjects received visual feedback of cursor position and reached to different targets randomly positioned at 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°, and 150° (Fig. 2*B*). A baseline block consisting of 50 trials, similar to the baseline block in *experiment 1*, followed familiarization trials. The baseline block was followed by a prelocalization block and a pregeneralization block. Next, the subjects experienced a rotation block of 400 trials, where a 30° counterclockwise abrupt visuomotor rotational perturbation was applied. The rotation block was followed by postlocalization and postgeneralization blocks. The environment then returned to a 0° rotation for the final 50 trials (washout block).

LOCALIZATION BLOCKS. In both the prelocalization and postlocalization blocks (50 trials each), subjects performed four reaching trials with the same feedback provided as in the baseline block. They were then asked to hold the robotic arm with their left hand on the fifth (localization) trial and locate where on the target arc (Fig. 1*B*) their right hand crossed in the immediately preceding trial (i.e., the 4th trial). This was repeated 10 times in each block for a total of 10 prelocalization trials and 10 postlocalization trials. Regardless of the type of feedback they received during the four trials, subjects in all groups could see the cursor on the localization trials.

GENERALIZATION BLOCKS. The pre- and postlocalization blocks were followed by pre- and postgeneralization blocks, respectively. In both the pre- and postgeneralization blocks (96 trials each), subjects reached to targets randomly positioned at 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°, and 150° (Fig. 2*B*). The frequencies of the target appearing at 90° and the target appearing at any other direction were, respectively, 32/96 and 8/96. Feedback was only provided when the subject reached to the central target at 90° and matched the feedback

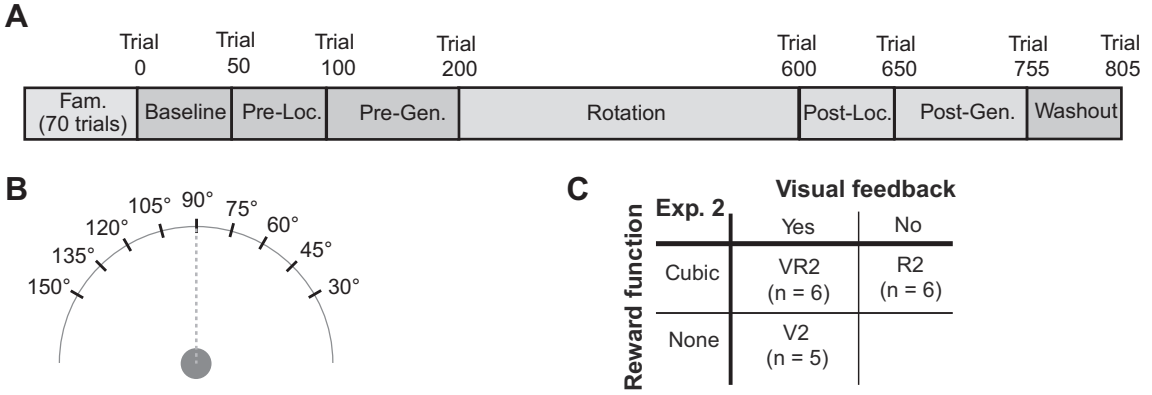

Fig. 2. *Experiment 2*: description. *A*: timeline of experimental blocks for Experiment 2. Fam, familiarization; Pre-Loc., prelocalization; Pre-Gen., pregeneralization; Post-Loc., postlocalization; Post-Gen., postgeneralization. *B*: target directions in the generalization blocks (all shown at once). *C*: design of experimental groups for *experiment 2*: a Control group (V2) that received only visual and no reward feedback, a group that received visual feedback and cubic reward feedback (VR2), and a group that received only cubic reward feedback (R2).

the subject had received during the baseline block. No feedback was provided in any of the groups when subjects reached to targets other than the central target at 90° (i.e., targets positioned at 30°, 45°, 60°, 75°, 105°, 120°, 135°, and 150°).

Experiment 2: groups. Subjects ($n = 17$) were divided into three groups (Fig. 2*C*), including a Control group (V2; $n = 5$) that received only visual and no reward feedback, a group that received visual feedback and cubic reward feedback (VR2; $n = 6$), and a group that received only cubic reward feedback (R2; $n = 6$).

Experiment 2: data analysis. In *experiment 2* we used the error and learning rate metrics to assess learning. We also introduced two additional metrics to quantify sensorimotor remapping (localization index, LI) and generalization (generalization index, GI) for *experiment 2*.

LOCALIZATION INDEX. At each localization trial, LI was defined as the angle between the lines connecting home to the points on the target arc where the left and the right hands crossed the target arc in two consecutive trials. Similar to the definition used to calculate error in *experiment 1*, a clockwise direction was taken as positive. This means that if the perceived hand position on the target arc at a localization trial was located on the right side of the actual crossing point in the trial right before, the LI should be positive. For each subject, LI was calculated on each localization trial and then averaged across all 10 localization trials separately for the pre- and postlocalization blocks. To assess sensorimotor remapping in different experimental groups, LI at the pre- and postlocalization blocks were compared across all groups. To this end, a three-way ANOVA ($2 \times 2 \times 2$) was carried out where the main effects and interactions of reward landscape, visual feedback, and block were evaluated (reward and visual feedback as between-subjects factors and block as a within-subjects factor).

GENERALIZATION INDEX. For each target direction, GI was defined as the change in the initial reaching angle from the pre- to the postgeneralization block. The line connecting home to the target position was used as the reference to calculate the initial reaching angle for each target direction. Clockwise directions were taken as positive. For each subject, GI was calculated at each target direction and was then averaged across all trials in which that target appeared and separately at the pre- and postgeneralization blocks. To compare generalization across groups, a four-way ANOVA ($2 \times 2 \times 9 \times 2$) was carried out where the main effects and interactions of reward feedback, visual feedback, target direction, and block on GI were evaluated (reward and visual feedback, and target direction as between-subjects factors and block as a within-subjects factor).

Model Predictions

A temporal difference (TD) learning model (Sutton and Barto 1998) was also used to determine the feasibility of learning an abrupt visuomotor rotation by reward feedback alone and to investigate the mechanisms underlying learning in the different reward landscapes. The TD modeling framework has been very successful in approximating reward-based learning in the brain (Schultz 2013). There is, however, no consensus about which TD modeling architecture best represents reward-based learning. The controversy essentially revolves around whether the brain learns the policy (O'Doherty 2004) or the Q values (Roesch et al. 2007). In this study we decided to use a standard Q-learning modeling architecture with a softmax as the action-selection algorithm. The action-value function, $Q(s,a)$, was updated at each trial using the following formulation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[\gamma + \gamma \max_a Q(s',a) - Q(s,a)\right], \quad (4)$$

where α and γ are, respectively, the learning and the discount rate. The second term on the right side of *Eq. 4* is the reward prediction error. For each selected action a at state s, the value of reward was calculated using *Eq. 1*, with θ having the same value as a. Only two states were considered: the initial (s) and the terminal (s') states (there were no intermediate states). The action a was defined to be the initial reaching angle selected by the subject at the initial state s. For modeling purpose, the action space A was discretized into angles from $-90°$ to $90°$ with a step size of $0.1°$ (in total, 1,801 possible actions). By choosing action a at each trial, the policy π was updated using the Gibbs softmax function as follows:

$$\pi(s,a) \leftarrow \frac{e^{Q(s,a)/\tau}}{\sum_{b=1}^{A(s)} e^{Q(s,a)/\tau}}, \quad (5)$$

where τ is constant and called the "temperature." Similar to the experimental protocol (Fig. 1*D*), the rotational perturbation was abruptly introduced after the 50th trial and was suddenly removed after the 500th trial. Because the general trends predicted by the model were not sensitive to the selected values for parameters and the initial conditions, arbitrary values were used for modeling simulations ($n = 2{,}000$).

The early and the late phases were defined as the first 10 and the last 10 trials in each test block. For each reward landscape, error at the early learning (EL) and early washout (EW) phases was compared with that at, respectively, the late learning (LL) and late washout (LW) phases using paired *t*-tests. Confidence intervals for the mean error were used to compare the learning rate (in the rotation block) between the two reward landscapes. Three different intervals for comparison were defined as for the experimental data.

Statistical Analysis

Unless otherwise noted, SPSS (version 22; IBM SPSS Statistics) was used for statistical analyses. The threshold of statistical significance was set at $\alpha = 0.05$. Post hoc tests applied a Bonferroni correction. Whenever the assumption of sphericity was violated, we applied a Greenhouse-Geisser correction on the degrees of freedom. For all statistical comparisons, P values are reported up to three significant digits, except for P values <0.001. Whenever mean values are given, SE is also presented (means ± SE).

RESULTS

Experiment 1

Reward alone can produce visuomotor learning. We first examined the question of whether it was possible to learn to compensate for an abrupt visuomotor perturbation without visual feedback and by using only reward feedback. Performance of the two groups that received only reward feedback (R_L and R_C) is shown in Fig. 3, *A* and *C*. Despite the fact that the reward landscape was positive definite, and therefore ambiguous with respect to the direction of error, the subjects in the R_L and R_C groups were able to compensate for the perturbation about as well as the subjects that received visual feedback (V group). The ANOVA revealed a main effect of phase on error [$F(2.969,121.718) = 319.569$, $P < 0.001$]. All groups significantly reduced error ($P < 0.001$; Fig. 3, *A* and *B*) and increased trial score R ($P < 0.001$; Fig. 3, *C* and *D*) from the EL to the LL phase. Learning was also quickly washed out in all groups in that there was a significant drop in error ($P < 0.001$) from the EW phase to the LW phase (Fig. 3, *A* and *B*). Furthermore, the lack of visual feedback in the No-Vision groups did not affect the extent of error reduction. Whereas the ANOVA showed a main effect of visual feedback [$F(1,41) = 7.251$, $P = 0.010$], all groups exhibited comparable error at the start of learning (EL phase, $P = 0.498$) and learned the task to a similar extent (LL phase, $P = 0.451$). There was a visual

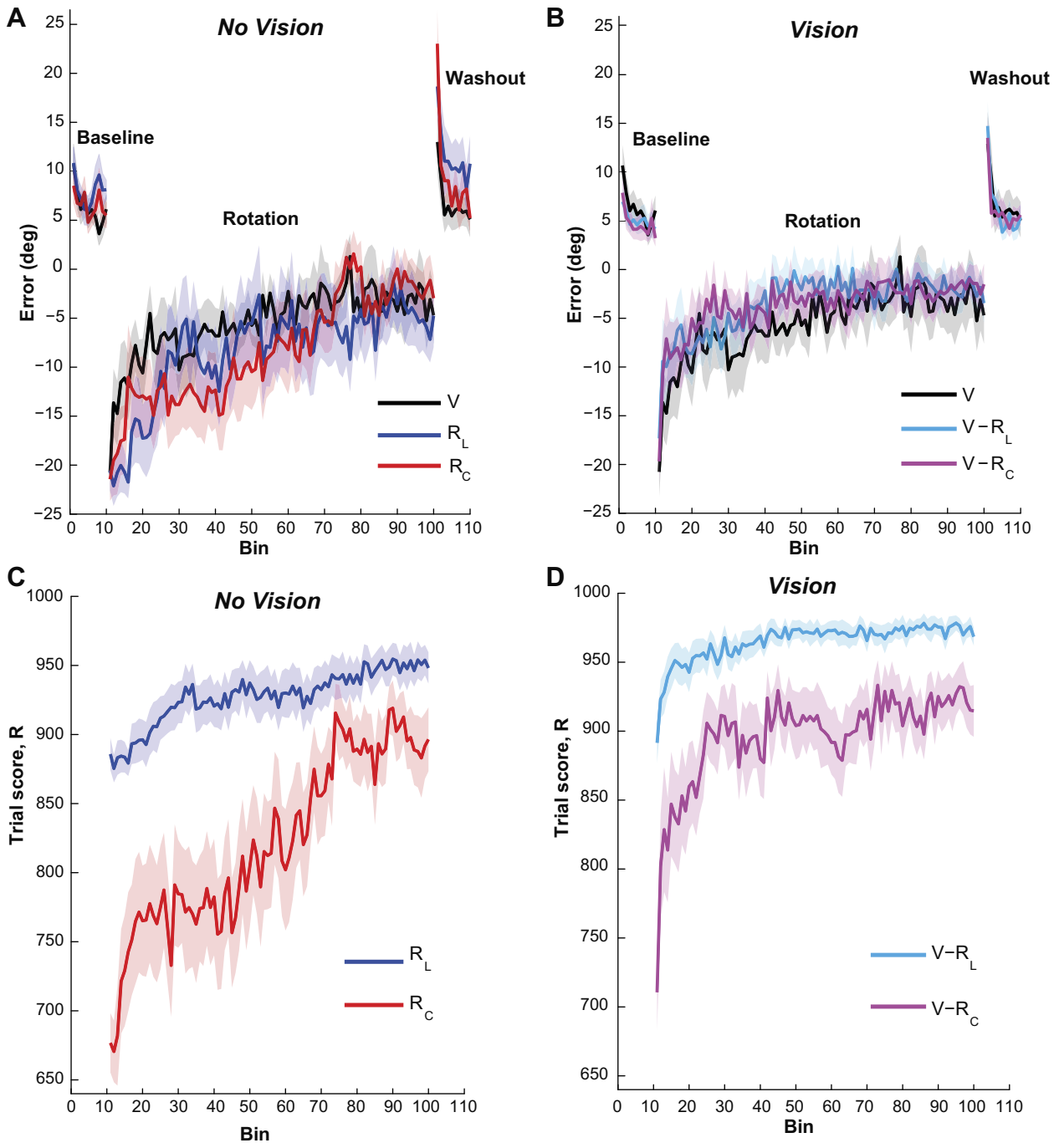

Fig. 3. *Experiment 1*: error. *A* and *B*: mean and SE of error (θ, degrees) vs. bin number for No-Vision (plus Control) groups (*A*) and Vision groups (*B*) during baseline, rotation, and washout. *C* and *D*: mean and SE of trial score (R; *Eq. 1*) vs. bin number for No-Vision groups (*C*) and Vision-Reward groups (*D*) in the rotation block. Error and trial score data have been averaged across all subjects in each group. A bin is defined as 5 consecutive trials.

feedback × phase interaction effect that resulted from error in the EW phase being significantly larger for the R_C group compared with the V ($P = 0.008$), V-R_L ($P = 0.024$), and V-R_C ($P = 0.003$) groups [$F(2.969,121.718) = 5.518$, $P = 0.001$]. The variability results paint a similar picture. Whereas there was a main effect of visual feedback [$F(1,41) = 11.552$, $P = 0.002$], no significant differences were found between the No-Vision groups and the Control group ($P = 0.324$ and $P = 0.560$ for R_L and R_C, respectively; Fig. 4). Additionally, the choice of the reward landscape did not affect error or variability, because the ANOVAs revealed no main effects of reward [error: $F(2,41) = 0.050$, $P = 0.952$; variability: $F(2,41) = 0.751$, $P = 0.478$].

Although subjects learned to a similar extent without vision, they learned more slowly (Fig. 3*A*). There was a high degree of variance across subjects (Fig. 5). Subjects roughly fell into one of three categories: fast-, slow-, and medium-latency learners. One of each is shown in Fig. 5 for the R_L and R_C groups. Because of this variability, the learning curves for the No-Vision groups were not well described by an exponential

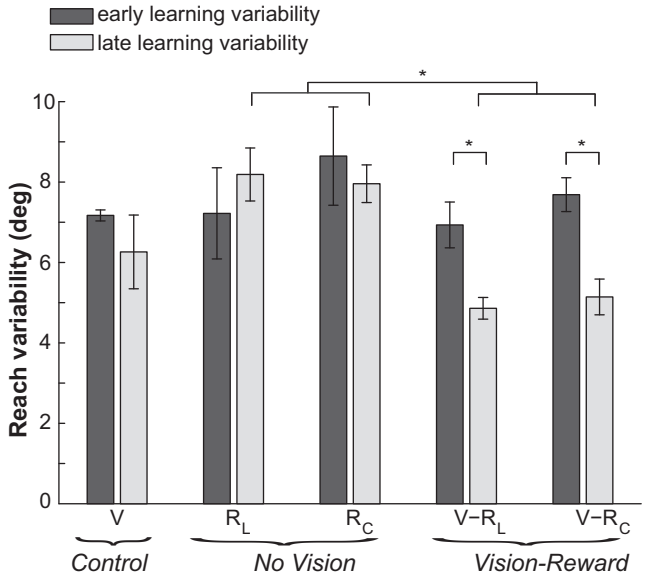

Fig. 4. *Experiment 1*: variability. Data indicate early leaning variability (first 100 trials) and late learning variability (last 100 trials) in the rotation block for each group. Bars represent mean variability across subjects. Error bars represent SE. $*P < 0.05$, statistically significant difference.

function. Therefore, we compared the mean error to estimate learning rates at different intervals. Regardless of the selected interval for comparison, mean error for the No-Vision groups was significantly greater than for the Control group, indicating a significantly faster learning rate in the Control group [1st interval (first 50 trials): -17.370 ± 0.420 vs. -12.160 ± 0.6530 (No-Vision vs. Control); 2nd interval (*trials 51–200*): -10.907 ± 0.205 vs. -6.809 ± 0.227; 3rd interval (*trials 201–450*): -4.884 ± 0.166 vs. -3.294 ± 0.183; $P < 0.001$ for all 3 comparisons].

Interestingly, the reward landscape seemed to affect learning rate, and the effect was dependent on the phase of learning.

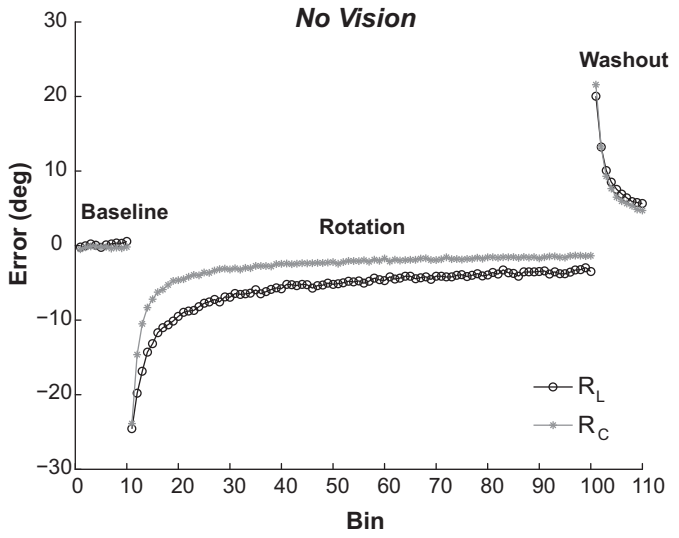

Fig. 6. Model-predicted error vs. bin number is shown for No-Vision conditions at different test blocks.

Initially, learning was significantly faster for the cubic reward landscape (1st interval, R_C: -15.676 ± 0.576, R_L: -19.059 ± 0.415, $P < 0.001$), but later, learning was significantly faster for the linear reward landscape (2nd interval, R_C: -12.460 ± 0.215, R_L: -9.353 ± 0.298, $P < 0.001$). Incorporating the later trials into the comparison demonstrated that learning was again faster for the cubic reward landscape (3rd interval: R_C: -3.876 ± 0.237, R_L: -5.515 ± 0.166, $P < 0.001$).

Similar to the experimental findings for the No-Vision groups, model predictions (Fig. 6) showed learning with reward feedback alone in that there was a significant decrease in error from EL to LL and also from EW to LW (paired *t*-test, $P < 0.001$). Also corroborating the experimental results, the two reward landscapes yielded the same learning level, because the mean errors at the LL phase are numerically close

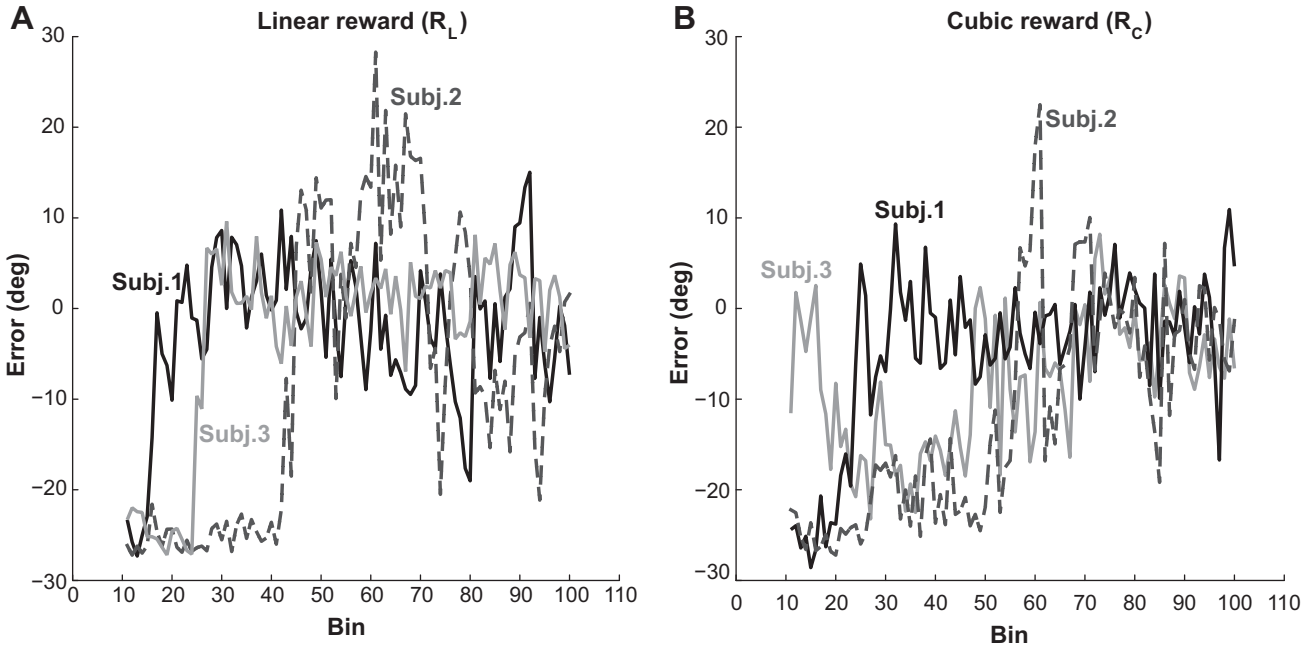

Fig. 5. *Experiment 1*: representative subjects. *A* and *B*: error (θ, degrees) vs. bin number during the rotation block for 3 representative subjects in the R_L (*A*) and R_C group (*B*). Subjects roughly fell into 1 of 3 categories: fast-, slow-, and medium-latency learners.

($-1.7°$ for R_C vs. $-3.4°$ for R_L). However, the model did not explain the experimentally observed learning rates. The model predicted a faster learning rate for the cubic reward landscape, throughout the rotation block, regardless of the selected interval for comparison. This was comparable to the experimental results only for the 1st interval and 3rd interval, but not for the 2nd interval, where the Linear group outperformed the Cubic group.

Reward accelerates learning when paired with visual feedback. We found that learning was faster when both reward and sensory feedback were provided, compared with only sensory feedback (Fig. 7). Although the extent of error reduction was similar, the average learning rate obtained in the Vision-Reward groups was significantly faster than the learning rate in the Control group ($c = 0.457 \pm 0.063$ and $c = 0.212 \pm 0.063$, respectively, $P = 0.047$; Fig. 7*B*). To provide further confirmation of these results, we also fit exponential curves to the average error curves in each group across the first 50 learning trials. Here, as well, the learning rates for two groups that received both visual and reward feedback (V-R_L: $c = 0.451 \pm 0.103$, V-R_C: $c = 0.287 \pm 0.051$) were significantly faster than the learning rate for the Control group (V: $c = 0.117 \pm 0.044$, 95% confidence interval; Fig. 7*D*).

The exponential fits to the learning curves of individual subjects confirmed that there was no difference between groups in a ($P = 0.700$) and b ($P = 0.132$) parameters, which together represent the initial and final learning levels for the 50-trial

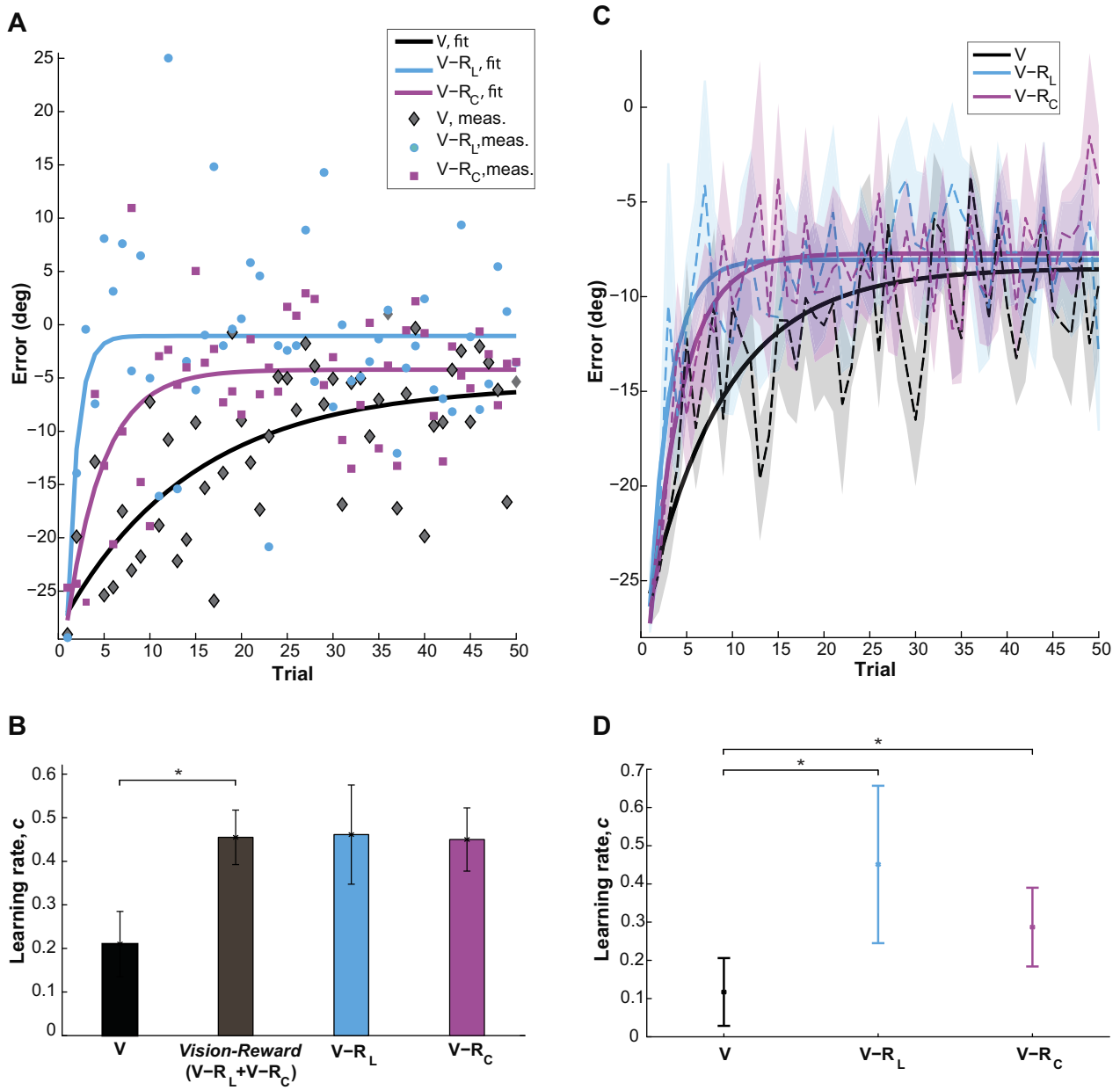

Fig. 7. *Experiment 1*: learning rate. Exponential fit to the error data in the first 50 trials of the rotation block. *A*: fit (solid lines) vs. measured error (dots) for a representative subject from each Vision group (V, V-R_L, and V-R_C). *B*: mean and SE of the learning rate c for fits to individual subjects in different groups. *C*: fit to the mean error in each group (solid lines) and mean error across all subjects in each group (dashed lines). Shaded regions represent SE. *D*: mean and 95% confidence intervals of the learning rate c for each Vision group. *$P < 0.05$, statistically significant difference.

epoch. Thus any differences in the learning rate (c; *Eq. 2*) would be sufficient to demonstrate differences between groups. Similarly, the confidence interval analysis performed on the fits to the average error curves indicated there were no significant differences in the value of coefficients a and b between the three groups.

Combining reward and visual feedback also led to greater reductions in variability. Variability significantly decreased with learning only for the groups that received both visual and reward feedback (V-R_L: $P = 0.013$, V-R_C: $P = 0.002$; Fig. 4). By late learning, variability in these two groups was significantly less than when reward was the only source of feedback (R_L vs. V-R_L: $P = 0.001$, R_L vs. V-R_C: $P = 0.002$, R_C vs. V-R_L: $P = 0.003$, R_C vs. V-R_C: $P = 0.005$).

Possible Learning Mechanisms

In *experiment 2*, subjects also performed generalization and localization trials, with which we could examine the underlying learning mechanisms. Results of *experiment 2* generally reproduce the findings of *experiment 1*.

Reward alone can produce visuomotor learning. Similar to *experiment 1*, subjects in the *experiment 2* Reward (R2) group were able to compensate for the perturbation as well as the subjects that received visual feedback (V2 and VR2 groups). The ANOVA revealed a main effect of phase on error [$F(3.590,50.258) = 182.685$, $P < 0.001$]. All groups significantly reduced error ($P < 0.001$; Fig. 8*A*) and increased trial score R ($P < 0.001$) from the EL to the LL phase. Learning was quickly washed out in all groups in that there was a significant drop in error (V2: $P < 0.001$, VR2: $P = 0.009$, R: $P = 0.001$) from the EW phase to the LW phase (Fig. 8*A*). The ANOVA showed no main effect of either visual feedback [$F(1,14) = 2.179$, $P = 0.162$] or reward landscape [$F(1,14) = 0.185$, $P = 0.673$] on error. All groups learned the task to a similar extent (LL phase, $P = 0.158$). There was a visual feedback × phase interaction effect [$F(3.590,50.285) = 4.030$, $P = 0.008$] that resulted from error in the EL phase being significantly larger ($P = 0.018$) for the R2 group compared with the V2 group.

Reward accelerates learning when paired with visual feedback. Similar to *experiment 1*, results from *experiment 2* also showed that combining reward and sensory feedback resulted in faster learning. To quantify learning rate, we first fit exponential curves to the average error curves in each group across the first 50 learning trials (Fig. 8*B*). However, with the use of the method presented in *Eq. 3*, there was a significant difference in parameter a (*Eq. 2*) between the Control (V2: $a = -11.599 \pm 0.435$) and either the Reward (R2: $a = -3.140 \pm 1.768$) or the Vision-Reward (VR2: $a = -3.447 \pm 0.708$) groups. Since parameter c could not be directly used to compare learning rate between the Control and the other two groups, we turned to the alternative method of comparing mean error (Fig. 8*B*). This comparison revealed a significantly larger ($P < 0.001$) error for the Control group (V2: -12.045 ± 0.473) compared with the Vision-Reward group (VR2: -5.683 ± 0.851), indicating a faster learning rate for the latter. Additionally, both the exponential fit (VR2: $c = 0.213 \pm 0.053$, R2: $c = 0.067 \pm 0.019$) and comparison of mean error (Fig. 8*B*) confirmed a significantly faster learning rate for the Vision-Reward (VR2) group compared with the Reward (R2) group.

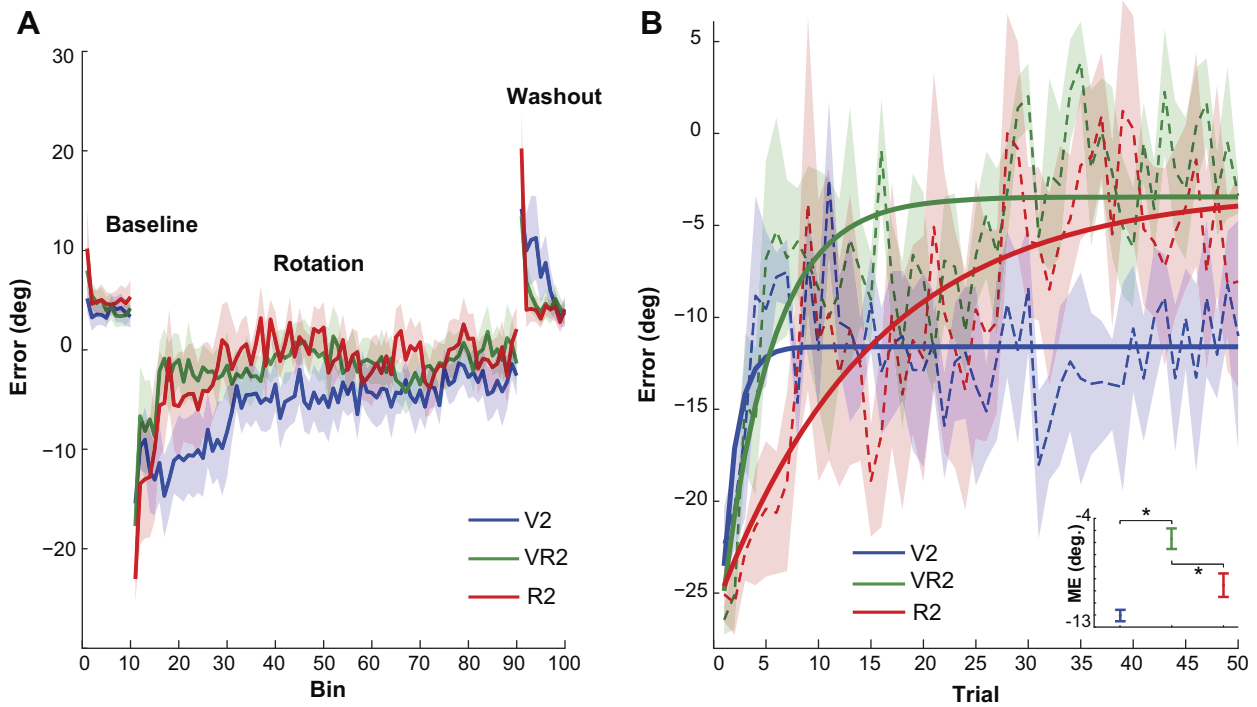

Fig. 8. *Experiment 2*: error and learning rate. *A*: mean and SE of the error (θ, degrees) vs. bin number for the 3 experimental groups during baseline, rotation, and washout. Error data have been averaged across all subjects in each group. Every 5 measured trials are defined as a bin. *B*: exponential fit to the mean error data in the first 50 trials of the learning block in each group (solid lines) vs. mean error curves across all subjects in each group (dashed lines). *Inset* at *bottom right* represents mean error (ME) across all subjects and in the first 50 trials of the learning block for the 3 measured groups. Error bars and shaded regions represent SE. *$P < 0.05$, statistically significant difference between groups.

Combining reward and sensory feedback does not compromise sensorimotor remapping. Subjects' accuracy in the estimation of hand position (Fig. 9*A*) significantly decreased from the pre- to postlocalization block for the groups who received visual feedback of the cursor (V2: $P < 0.001$, VR2: $P = 0.001$), indicating that learning led to a similar degree of remapping in both groups. However, it did not significantly change in the Reward (R2) group ($P = 0.163$). The ANOVA revealed a main effect of block [$F(1,14) = 85.882$, $P < 0.001$], visual feedback [$F_{(1,14)} = 26.295$, $P < 0.001$], and a block × visual feedback interaction [$F(1,14) = 14.578$, $P = 0.002$] on LI. Regarding the reward feedback, neither a main effect of reward on LI [$F(1,14) = 0.946$, $P = 0.347$] nor a block × reward interaction [$F(1,14) = 0.656$, $P = 0.432$] was observed. Although no significant difference in LI was observed between Vision groups at any of the pre- and/or postlocalization blocks, both groups (V2 and VR2) showed a significant difference from the Reward (R2) group at both the prelocalization (V2 vs. R: $P = 0.029$, VR2 vs R2: $P = 0.046$) and post- localization blocks (V2 vs R: $P < 0.001$, VR2 vs. R2: $P = 0.001$).

Learning generalizes to nearby targets in all groups. An interesting observation in *experiment 2* was that all subjects, including the ones who received only reward feedback, could generalize learning to untrained, nearby targets (Fig. 9*B*). The ANOVA revealed a main effect of block on GI [$F(1,126) = 425.376$, $P < 0.001$]. Initial reaching angle increased from the pre- to the postgeneralization block (Fig. 9*B*), but no main effect of visual feedback [$F(1,126) = 1.073$, $P = 0.302$] or reward feedback [$F(1,126) = 0.694$, $P = 0.406$] was observed. Post hoc analyses revealed that the change in angle was significant in all cases except for R2 group reaching to target angle 105° ($P = 0.117$) and for V2 group reaching to target angles 135° ($P = 0.253$) and 150° ($P = 0.311$). A main effect of target direction on GI [$F(8,126) = 20.580$, $P < 0.001$] and a block × target direction interaction [$F(8,126) = 4.056$, $P < 0.001$] was also observed. Post hoc analysis showed that the mean difference between the two extreme target angles on the opposite ends of the rotational perturbation and the center target were significant (30° vs. 90°: $P < 0.001$; 45° vs. 90°: $P < 0.001$; 60° vs. 90°: $P < 0.001$; 105° vs. 90°: $P < 0.001$; 120° vs. 90°: $P = 0.001$), but not those in the same direction as the rotation (135° vs. 90°: $P = 1.000$; 150° vs. 90°: $P = 1.000$).

DISCUSSION

In this study we sought to examine the role of reward in motor learning. We asked whether it is possible to learn an abrupt visuomotor rotation with reward feedback alone (Q1) and whether the learning process can be modulated by combining both reward and sensory feedback (Q2). Together our findings indicate that it is possible to learn an abrupt visuomotor rotation using reward feedback alone and that the combination of reward and sensory feedback accelerates learning compared with either form of feedback alone.

All groups, but most importantly, even the groups that received only reward feedback, were able to learn the abrupt visuomotor rotation task as the error significantly decreased at the end of the learning block compared with the error at the movement onset. Subjects who received only reward feedback could also learn the task to the same extent as those who received either visual feedback alone or a combination of reward and error feedback. This means that not only can healthy adults learn from binary reward when perturbations were small and introduced gradually, as was shown by Izawa and Shadmehr (2011), but they also are able to learn from continuous reward feedback alone when perturbations are large and introduced abruptly in a visuomotor rotation task.

The breadth of the reward landscapes in our experiment was an important feature that could possibly explain the success of learning from reward alone in the presence of abrupt perturbation. Whereas the reward landscapes in this study could cover the regions far beyond 30°, the region that yielded binary reward in the study by Izawa and Shadmehr (2011) was small and limited to angles around the target. They applied a gradual perturbation in their experiment to make sure that the subject could receive some reward at least some of the time. A reward landscape that is not only continuous but also broad could

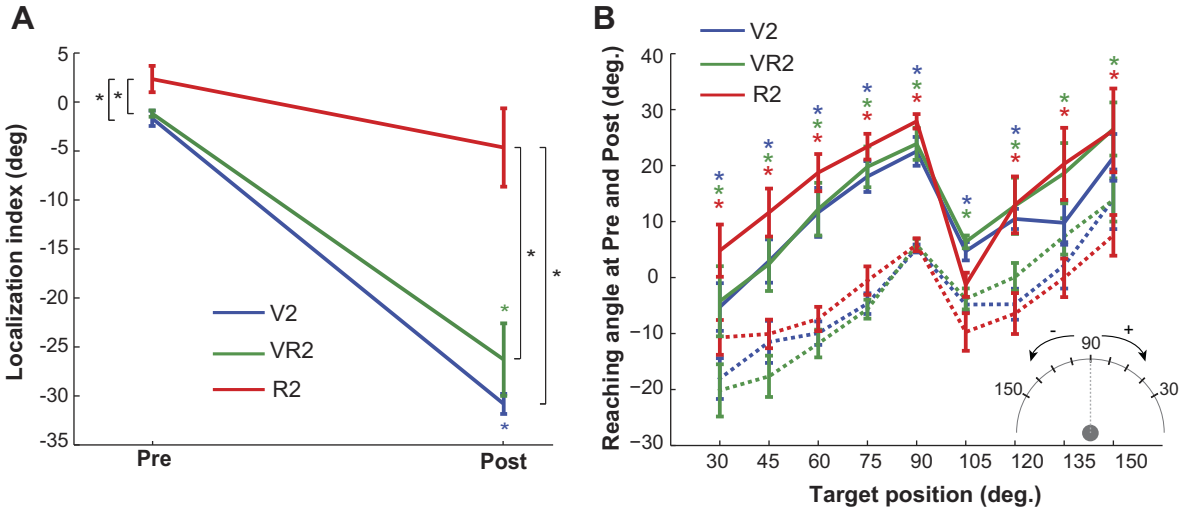

Fig. 9. *Experiment 2*: localization and generalization. *A*: localization index (LI) in the pre- and postlocalization blocks for all groups. Black asterisks indicate differences between groups ($P < 0.05$). Colored asterisks indicate group-specific decrease in LI from pre- to postlocalization. *B*: initial reaching angle in the pre- and postgeneralization blocks for all groups (dashed and solid lines, respectively). Asterisks indicate an increase in generalization index from pre- to postgeneralization. Error bars represent SE.

provide subjects with informative feedback even in the early adaptation when the errors are large. The reward gradient may also explain why we observed a broad generalization pattern for the Reward group (similar to the Vision groups), whereas others have not (Izawa and Shadmehr 2011). Thus it appears that a gradient of reward feedback presents a happy information medium between binary reward and visual feedback.

Despite the fact that it was possible to learn the visuomotor rotation task with reward alone, there were significant differences with this form of learning compared with the condition in which visual feedback was also present. Notably, the learning rate was slower. In some sense, this is surprising, since error and reward were strongly correlated as long as subjects did not overshoot the target. Early in the learning block, subjects rarely overshot the target in any of the groups. Yet even early in the learning block, learning proceeded at a slower rate in the No-Vision groups. Learning was slower likely due to the reduced amount of information provided. In the Vision groups, the signed error provided them with a clear indication of how to correct on the next movement. In contrast, in the No-Vision groups, error was unsigned, so the direction in which they should correct was not clear based on the feedback provided in the previous trial.

Adding the reward feedback on top of the visual feedback accelerated the learning rate. Endpoint variability was significantly reduced with learning only for subjects receiving both reward and visual feedback, even though subjects in other groups started to learn the task with comparable variability. The endpoint variability was also smaller for the Vision-Reward groups compared with the No-Vision groups. Combining the reward and the sensory feedback significantly accelerated learning and decreased the endpoint reach variability compared with the vision- or reward-alone conditions, likely because it is more informative than vision or reward feedback alone.

Learning Mechanisms

An important question at this junction is, what representational change is occurring as learning progresses in the No-Vision groups and the Vision-Reward groups, and how does it compare to the Control group? Although subjects in the No-Vision groups could learn the task to the same extent, they learned it more slowly and exhibited more variability on completion of the learning block. Similarly, although the Vision-Reward groups also learned the task to same extent, they learned it faster and with less final variability than the Control group. In the following paragraph, we discuss potential learning mechanisms that may be contributing individually or in combination to learning in one or more of the groups. One possibility is that subjects learned an explicit aiming strategy, i.e., "aim 30° clockwise to compensate for the 30° counterclockwise rotation." In studies where subjects have been verbally instructed by the experimenters to use an explicit aiming strategy, error will gradually drift in the opposite direction, due to an aiming error (Mazzoni and Krakauer 2006). However, it was recently shown that explicit strategies contribute to the learning process even in the absence of experimenter instructions (Taylor et al. 2014). This is also analogous to the proposal of Izawa and Shadmehr (2011) that subjects are updating action selection on the basis of the reward prediction error, the difference between the reward predicted and that realized. Taking all the above together, an explicit strategy or action selection may be driving all or a portion of the learning observed. A second alternative is that, as in traditional visuomotor rotation tasks where visual feedback is provided, subjects are updating a sensorimotor map describing the relationship between hand and cursor location. An impressive body of literature has demonstrated that this process is driven by sensory prediction errors and is cerebellum dependent. Finally, the presence of a reward gradient allows for a third option. It is possible that the subjects learned a new representation of the reward structure, i.e., a mapping describing the relationship between arm movement and reward. This alternative would fall somewhere between the first and second options mentioned above. A reward structure is more complex than action selection yet does not require a change in the sensorimotor mapping between the hand and the cursor.

Izawa and Shadmehr (2011) used both generalization and localization experiments to distinguish between an update of a sensorimotor mapping and an action policy (first and second options above). Specifically, learning from reward should lead to an update of an action policy, and therefore learning is local and does not generalize to nearby targets. This process does not alter the sensorimotor mapping as assessed with a localization task. In contrast, learning from sensory prediction error updates the sensorimotor representation, which thus leads to both generalization of learning to nearby targets and improved performance in the localization task.

Our localization results provide strong evidence that learning from reward feedback does not lead to an update of the sensorimotor map of the relationship between hand and cursor location. Subjects in the No-Vision groups were amazingly accurate when asked to localize the position of their hand on the previous trial, as accurate as they were prior to exposure to the perturbation. If the presence of a reward gradient leads to updating of a representation of the reward structure (third option above), then we would predict that learning will generalize to nearby targets but will not affect localization. This third option, i.e., learning a mapping between arm movement and reward, is strongly supported by our findings. Subjects in the No-Vision groups could generalize to nearby targets and could generalize to the same extent as those in the Control and Vision-Reward groups.

Using this approach, we can also attempt to understand the mechanisms underlying the increased rate of learning when sensory and reward feedback are combined in the Vision-Reward groups. One question, for example, is whether the superposition of reward changes the learning rate of the sensorimotor mapping directly (i.e., multiplicative combination) or modifies a reward-based learning component that is combined additively. Izawa and Shadmehr (2011) developed a learning model that additively combined the changes dictated by the reward and sensory prediction error, but they did not explicitly compare learning rates across groups. It is also difficult to relate their findings to the present study, since they introduced the perturbation gradually and provided only binary reward feedback. Recently, Taylor et al. (2014) demonstrated, in an experiment where the visuomotor rotation was introduced abruptly, that the quality of visual feedback alters the relative contributions of explicit and implicit strategies but not their time course. If the superposition of reward and error led to an

increased contribution of explicit processes compared with error alone, then we would expect this to compromise sensorimotor remapping and we would see a reduction in sensorimotor remapping (i.e., a reduction in implicit learning). In contrast, our localization results demonstrate that the degree of remapping was similar between the Vision-Reward and Control groups. This suggests that the superposition of reward and error alters the learning rate of the sensorimotor mapping directly. A potential mechanism for the faster rate of internal model learning comes from the predictions of models that rely on Bayesian inference to modulate the learning rate. These models predict that the rate of internal model learning will be reduced when the uncertainty of the sensory feedback is increased, and their predictions have been experimentally confirmed (Burge et al. 2008; Wei and Kording 2010). Conversely, these models also predict that a reduction in the uncertainty of sensory feedback will increase the rate of learning, but there has been no experimental confirmation. One hypothesis is that the superposition of reward on the visual feedback of the error increases one's certainty in the sensory feedback, and thus increases the internal model learning rate. However, an alternative explanation is that the explicit and implicit processes are not additive and interact in a complex manner. Ultimately, we must wait for future research to determine the answer.

In recent years, evidence has emerged that provides support for the hypothesis we put forward regarding the update of an internal representation of the reward structure. Reinforcement learning theory has proven a powerful framework to understand the myriad of processes underlying reward-based learning. It has gained much traction within the neuroscience community because the behavior of dopaminergic neurons during reward-based learning tasks closely approximates reward prediction error. Reward prediction error is a critical learning signal in one of the most successful forms of reward-based learning models, temporal difference (TD) models (see Schultz 2013 for review). Usually, a simple TD learning rule is applied that updates action policies on the basis of momentary events. However, the community has recently begun to appreciate the importance of the underlying reward structure (Nakahara and Hikosaka 2012). Here, we use the word "structure" to describe the general context of the task or state representation that may include, among other things, a history of past events. Structural reinforcement learning models do a good job approximating human performance in a multi-armed bandit task, a reward-based learning task (Acuña and Schrater 2010). Perhaps most importantly, neurophysiological data confirm that dopaminergic activity more closely approximates the learning signal in a TD model that has knowledge of the reward structure, better than a traditional TD model with no knowledge of the reward structure (Nakahara et al. 2004). In the present experiment, subjects may construct a representation of the reward structure over multiple trials and use that information to improve reward prediction and action selection.

Effects of Different Reward Landscapes

The experimental results (Fig. 3, *A* and *C*) reveal a faster learning rate in the Cubic group at movement onset, which is surpassed by the linear reward landscape for the next 150 trials, until the Cubic group again prevails by the end of the learning block. Despite these differences, one should be cautious in an interpretation between the two landscapes because of the high degree of variance across subjects. No significant difference in terms of adaptation-error correlation was also found between the two landscapes. However, if there are indeed differences, one possible explanation for the faster rate of learning in the cubic reward landscape is the "anchoring effect (Furnham and Boo 2011). According to the anchoring effect, decisions can be biased toward a reference, an initially presented value to the subject. It is likely that our subjects were biased toward the maximum value of reward ($R \sim 1{,}000$). With abrupt application of the rotational perturbation, the starting value of reward ($R \sim 870$) presented to the Linear group was not considerably different from the reference value, whereas it was noticeably smaller ($R \sim 650$) than the reference value in the Cubic group. Subjects in the Cubic group thus could have had high motivation to explore and quickly increase their gain at movement onset (the first 50 trials). The same reason may explain why the learning rate was suddenly decreased after about 50 trials; subjects seemed to be satisfied with their performance ($R \sim 750$, much better than $R \sim 650$ on initial exposure). There is no clear explanation why it took about 150 trials for the subjects in the Cubic group to start to explore the environment again and maximize their reward. It would be interesting to examine to what extent cognitive biases are responsible for this quasi-stepwise behavior.

In the Vision-Reward groups, we could not discern any differences between the two reward landscapes. This may be attributable to the overwhelming influence of sensory feedback in driving the learning process. Indeed, the shape of the learning curves are well fit initially with a single-rate exponential, similar to the data reported in multiple standard visuomotor rotation tasks. The single target task may have also been too easy; multiple targets may slow the rate of adaptation and elucidate difference therein. It would be useful to develop a model for combined reward and sensory feedback that would allow us to further investigate, in a principled manner, whether manipulating the structure of the reward landscape can modulate the time course of the motor learning process.

Reinforcement Learning Model

The model used in this study provided basic proof-of-concept of learning from reward feedback alone and motivated the choice of reward landscape. However, the model and the experiment yield different conclusions regarding which reward landscape can lead to a faster learning rate throughout the learning block. Initially, because of the high degree of variance across different subjects (Fig. 5), the average learning curve may not be reflective of the individual learning behavior and thus may not be compared with the model estimation. If we assume that averaging can reflect the general behavior of the reward-based learning, the differences between the model and experiment could be explained in different ways. First, no intermediate state was considered in our model despite there being an infinite number of intermediate states in the experiment. This modeling assumption was made by considering that the reward given to the subjects depended only on the action selected at the first (or a few early) state(s) of the movement. The trial score (reward), however, provided no explicit information about "where" during the movement the subject may have been rewarded. Considering such uncertainty in the mod-

eling process may fill some gaps between the model prediction and the experimental observation. Another possible reason for the difference could be related to the mechanism underlying learning from reward feedback alone. Our choice of TD model is based on the assumption that such learning follows a model-free reinforcement learning process (Glaescher et al. 2010), i.e., learning from reward prediction error without the need of building a model of the environment. A model-based mechanism would, on the other hand, need to acquire a thorough knowledge about the environment. Future research is needed to examine possible computational mechanisms underlying reward-based learning.

Advocating for Reward in Motor Adaptation

We believe our results point to a fundamental role played by reward in the motor adaptation process. Great strides have been made in recent years in our understanding of the motor adaptation process, one of the foundational findings being that it is cerebellum dependent and is driven by sensory prediction error. A study by Shmuelof et al. (2012) suggests that although adaptation happens through a sensory error-based learning process, using (binary auditory) reward feedback during the asymptotic phase of adaptation can lead to long-term retention of the learned movement. Abe et al. (2011) also found that presence of reward enhances learning of a skill task. However, a number of studies have demonstrated that adaptation is disproportionately sensitive to error statistics and error size (Marko et al. 2012; Wei and Kording 2009). Significant advances have also been made in the last decade in decision neuroscience, which have led to a new appreciation of the influence of reward and their valuation on decision making. Movement also represents a decision-making process, influenced by rewards and penalties (Wolpert and Landy 2012). Previous findings from our laboratory support the idea that adaptation can be influenced by the error magnitude as well as its subjective value. Specifically, the cost or threat associated with an error can lead to a reduction in adaptation (Manista and Ahmed 2012; Trent and Ahmed 2013). Point rewards and penalties may represent an additional means with which to modulate subjective value, and ultimately the adaptation process.

Clinical Implications

Our results showed that a combination of reward and sensory feedback considerably improved motor learning in terms of learning rate in healthy adults. It is likely that both sensory and reward feedback drive learning when the two are combined. Reward-related learning is impaired in patients with neurodegenerative diseases that are associated with the malfunctioned dopamine transmission such as Parkinson's disease (Shohamy et al. 2005), Schizophrenia (Lau et al. 2013), and Huntington's disease (Chen et al. 2013). People with these same neurological disorders have shown little difficulty in motor adaptation to novel arm dynamics (Agostino et al. 1996; Smith and Shadmehr 2005) and to a visuomotor rotation (Bédard and Sanes 2011; Marinelli et al. 2009). On the other hand, patients with cerebral damage show deficits in motor adaptation from sensory feedback (Gibo et al. 2013). An interesting question for the future is whether both mechanisms participate in the learning process when the two types of feedback are combined. If so, to what extent and at which stages of learning would any of these mechanisms be in charge in healthy adults and in patients with neurodegenerative diseases?

Conclusions

We have shown that by abruptly introducing perturbations of significant size in a visuomotor learning task, it is possible to learn from reward feedback alone. We also have found that the combination of reward and sensory feedback accelerates learning and improves final performance. Learning from reward feedback could rely on the formation of a mapping between arm position and reward structure, because this type of learning does not alter the sensorimotor remapping but generalizes to the nearby targets. This study suggests that reward is a powerful tool that may be used either as a substitute or as a supplement to sensory feedback during motor learning, with the potential to improve current neurorehabilitation approaches.

ACKNOWLEDGMENTS

We thank R. Shadmehr for helpful comments on an earlier version of this manuscript. We also thank M. C. Trent for help with data collection.

GRANTS

This work was supported by Defense Advanced Research Projects Agency Young Faculty Award D12AP00253 and National Science Foundation Grants SES 1230933, SES 1352632, and CMMI 1200830 (to A. A. Ahmed).

DISCLOSURE

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

AUTHOR CONTRIBUTIONS

A.A.N. performed experiments; A.A.N. and A.A.A. analyzed data; A.A.N. and A.A.A. interpreted results of experiments; A.A.N. prepared figures; A.A.N. and A.A.A. drafted manuscript; A.A.N. and A.A.A. edited and revised manuscript; A.A.N. and A.A.A. approved final version of manuscript; A.A.A. conception and design of research.

REFERENCES

Abe M, Schambra H, Wassermann Eric M, Luckenbaugh D, Schweighofer N, Cohen LG. Reward improves long-term retention of a motor memory through induction of offline memory gains. *Curr Biol* 21: 557–562, 2011.

Acuña DE, Schrater P. Structure learning in human sequential decision-making. *PLoS Comput Biol* 6: e1001003, 2010.

Agostino R, Sanes JN, Hallett M. Motor skill learning in Parkinson's disease. *J Neurol Sci* 139: 218–226, 1996.

Ahmed AA, Wolpert DM. Transfer of dynamic learning across postures. *J Neurophysiol* 102: 2816–2824, 2009.

Bédard P, Sanes J. Basal ganglia-dependent processes in recalling learned visual-motor adaptations. *Exp Brain Res* 209: 385–393, 2011.

Burge J, Ernst MO, Banks MS. The statistical determinants of adaptation rate in human reaching. *J Vis* 8: 20 21–19, 2008.

Chen JY, Wang EA, Cepeda C, Levine MS. Dopamine imbalance in Huntington's disease: a mechanism for the lack of behavioral flexibility. *Front Neurosci* 7: 114, 2013.

Dam G, Kording K, Wei K. Credit assignment during movement reinforcement learning. *PLoS One* 8: e55352, 2013.

Dydewalle G. A new look at psychology of reinforcement in human learning. *Stud Psychol (Bratisl)* 24: 233–240, 1982.
Furnham A, Boo HC. A literature review of the anchoring effect. *J Socio Econ* 40: 35–42, 2011.
Gibo TL, Criscimagna-Hemminger SE, Okamura AM, Bastian AJ. Cerebellar motor learning: are environment dynamics more important than error size? *J Neurophysiol* 110: 322–333, 2013.
Glaescher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66: 585–595, 2010.
Glimcher PW, Camerer CF, Fehr E, Poldrack RA. *Neuroeconomics: Decision Making and the Brain*. London: Elsevier Academic, 2009.
Hoffman H, Theodorou E, Schaal S. Optimization strategies in human reinforcement learning. In: *Advances in Computational Motor Control VII*. Washington, DC: Am Soc Neurorehabil, 2008.
Huang HJ, Ahmed AA. Older adults learn less, but still reduce metabolic cost, during motor adaptation. *J Neurophysiol* 111: 135–144, 2013.
Huang HJ, Kram R, Ahmed AA. Reduction of metabolic cost during motor learning of arm reaching dynamics. *J Neurosci* 32: 2182–2190, 2012.
Izawa J, Shadmehr R. Learning from sensory and reward prediction errors during motor adaptation. *PLoS Comput Biol* 7: e1002012, 2011.
Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. *J Artif Intell Res* 4: 237–285, 1996.
Kormushev P, Calinon S, Caldwell D. Reinforcement learning in robotics: applications and real-world challenges. *Robotics* 2: 122–148, 2013.
Krakauer JW, Mazzoni P. Human sensorimotor learning: adaptation, skill, and beyond. *Curr Opin Neurobiol* 21: 636–644, 2011.
Krakauer JW, Pine ZM, Ghilardi MF, Ghez C. Learning of visuomotor transformations for vectorial planning of reaching trajectories. *J Neurosci* 20: 8916–8924, 2000.
Lau CI, Wang HC, Hsu JL, Liu ME. Does the dopamine hypothesis explain schizophrenia? *Rev Neurosci* 24: 389–400, 2013.
Manista GC, Ahmed AA. Stability limits modulate whole-body motor learning. *J Neurophysiol* 107: 1952–1961, 2012.
Manley H, Dayan P, Diedrichsen J. When money is not enough: awareness, success, and variability in motor learning. *PLoS One* 9: e86580, 2014.
Marinelli L, Crupi D, Di Rocco A, Bove M, Eidelberg D, Abbruzzese G, Ghilardi MF. Learning and consolidation of visuo-motor adaptation in Parkinson's disease. *Parkinsonism Relat Disord* 15: 6–11, 2009.
Marko MK, Haith AM, Harran MD, Shadmehr R. Sensitivity to prediction error in reach adaptation. *J Neurophysiol* 108: 1752–1763, 2012.
Mataric MJ. Reward functions for accelerated learning. In: *The 11th International Conference on Machine Learning*. New Brunswick, NJ: Morgan Kaufmann, 1994, p. 181–189.
Mazzoni P, Krakauer JW. An implicit plan overrides an explicit strategy during visuomotor adaptation. *J Neurosci* 26: 3642–3645, 2006.
Nakahara H, Hikosaka O. Learning to represent reward structure: a key to adapting to complex environments. *Neurosci Res* 74: 177–183, 2012.
Nakahara H, Itoh H, Kawagoe R, Takikawa Y, Hikosaka O. Dopamine neurons can represent context-dependent prediction error. *Neuron* 41: 269–280, 2004.
Niekum S, Spector L, Barto AG. Evolution of reward functions for reinforcement learning. In: *Genetic and Evolutionary Computation Conference*. Dublin, Ireland: ACM, 2011, p. 177–178.
Nikooyan AA, Zadpoor AA. Application of virtual environments to assessment of human motor learning during reaching movements. *Presence Teleop Virt* 18: 112–124, 2009.
O'Doherty JP. Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Curr Opin Neurobiol* 14: 769–776, 2004.
Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9: 97–113, 1971.
Palminteri S, Lebreton M, Worbe Y, Hartmann A, Lehericy S, Vidailhet M, Grabli D, Pessiglione M. Dopamine-dependent reinforcement of motor skill learning: evidence from Gilles de la Tourette syndrome. *Brain* 134: 2287–2301, 2011.
Roesch MR, Calu DJ, Schoenbaum G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci* 10: 1615–1624, 2007.
Schlerf JE, Galea JM, Bastian AJ, Celnik PA. Dynamic modulation of cerebellar excitability for abrupt, but not gradual, visuomotor adaptation. *J Neurosci* 32: 11610–11617, 2012.
Schultz W. Updating dopamine reward signals. *Curr Opin Neurobiol* 23: 229–238, 2013.
Shmuelof L, Huang VS, Haith AM, Delnicki RJ, Mazzoni P, Krakauer JW. Overcoming motor "forgetting" through reinforcement of learned actions. *J Neurosci* 32: 14617–14621, 2012.
Shohamy D, Myers CE, Grossman S, Sage J, Gluck MA. The role of dopamine in cognitive sequence learning: evidence from Parkinson's disease. *Behav Brain Res* 156: 191–199, 2005.
Smith MA, Shadmehr R. Intact ability to learn internal models of arm dynamics in Huntington's disease but not cerebellar degeneration. *J Neurophysiol* 93: 2809–2821, 2005.
Sutton RS, Barto AG. *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press, 1998.
Taylor JA, Krakauer JW, Ivry RB. Explicit and implicit contributions to learning in a sensorimotor adaptation task. *J Neurosci* 34: 3023–3032, 2014.
Trent MC, Ahmed AA. Learning from the value of your mistakes: evidence for a risk-sensitive process in movement adaptation. *Front Comput Neurosci* 7: 118, 2013.
Wei K, Kording K. Relevance of error: what drives motor adaptation? *J Neurophysiol* 101: 655–664, 2009.
Wei K, Kording K. Uncertainty of feedback and state estimation determines the speed of motor adaptation. *Front Comput Neurosci* 4: 11, 2010.
Wolfe R, Hanley J. If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *Can Med Assoc J* 166: 65–66, 2002.
Wolpert DM, Landy MS. Motor control is decision-making. *Curr Opin Neurobiol* 22: 996–1003, 2012.
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.