Exp Brain Res (2008) 190:475–491
DOI 10.1007/s00221-008-1489-3

RESEARCH ARTICLE

Performance differences in visually and internally guided continuous manual tracking movements

Benjamin A. Philip · Yanchun Wu · John P. Donoghue · Jerome N. Sanes

Received: 11 July 2007 / Accepted: 7 July 2008 / Published online: 23 July 2008

Abstract Control of familiar visually guided movements involves internal plans as well as visual and other online sensory information, though how visual and internal plans combine for reaching movements remain unclear. Traditional motor sequence learning tasks, such as the serial reaction time task, use stereotyped movements and measure only reaction time. Here, we used a continuous sequential reaching task comprised of naturalistic movements, in order to provide detailed kinematic performance measures. When we embedded pre-learned trajectories (those presumably having an internal plan) within similar but unpredictable movement sequences, participants performed the two kinds of movements with remarkable similarity, and position error alone could not reliably identify the epoch. For such embedded movements, performance during pre-learned sequences showed statistically significant but trivial decreases in measures of kinematic error, compared to performance during novel sequences. However, different sets of kinematic error variables changed significantly between learned and novel sequences for individual participants, suggesting that each participant used distinct motor strategies favoring different kinematic variables during each of the two movement types. Algorithms that incorporated multiple kinematic variables identified transitions between the two movement types well but imperfectly. Hidden Markov model classification differentiated learned and novel movements on single trials based on the above kinematic error variables with 82 ± 5% accuracy within 244 ± 696 ms, despite the limited extent of changes in those errors. These results suggest that the motor system can achieve markedly similar performance whether or not an internal plan is present, as only subtle changes arise from any difference between the neural substrates involved in those two conditions.

Keywords Motor control · Reaching · Human · Sequence learning · Hidden Markov models

B. A. Philip (✉) · J. P. Donoghue · J. N. Sanes
Department of Neuroscience, Warren Alpert Medical School of Brown University, Sidney Frank Hall, 185 Meeting Street, Box GL-N, Providence, RI 02912, USA
e-mail: Benjamin_Philip@brown.edu

Y. Wu
Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912, USA

Introduction

Voluntary arm movements can be guided by visual cues and internal plans, separately or in combination. In order to have value as feedback, information about the current position of the hand or arm must be compared to a planned position or arm configuration. Such information can only proceed from current sensory (e.g., visual) input or comparison to an internal 'plan' generated through learning. While the more commonly used term "internal model" (see, e.g., Shadmehr and Mussa-Ivaldi 1994; Miall and Jackson 2006) refers to a forward model of arm dynamics as modified by environmental influences, this "internal plan" instead refers to prior knowledge about a movement sequence or task sequence; that is, the additional information available only for a previously learned movement pattern. It may be a movement template, procedural model of muscle activations, or memorized sequence of cues or other external guides; each of these possibilities represents one possible way in which the nervous system might store or use this prior knowledge. Behavioral differences

between movements guided with and without this internal plan can provide information about the neural mechanisms of internally and sensory guided actions.

Studies attempting to differentiate between trained and novel (untrained) movement sequences have primarily examined simple, stereotyped movements such as the finger-tapping patterns used in the serial reaction time task (SRTT, e.g., Nissen and Bullemer 1987; for discussion see Robertson 2007). These studies have been mainly motivated to understand the differences between explicit (declarative) and implicit (procedural) representations of learned sequences. In contrast, motor sequence learning has not been extensively studied in manual tracking movements that occur in many ecologically relevant situations such as driving a car or printing letters. For both of these movement types, the presence of an internal plan marks the distinction between learned and unlearned movements; performance improvements for practiced sequences over novel sequences reveal the use of such an internal plan in addition to on-line sensory information for control. These internal plans of procedurally learned sequences appear to depend upon element-to-element associations in primary motor cortex and basal ganglia, with temporal integration in supplementary motor area and pre-supplementary motor area, and rehearsal in prefrontal cortex as explicit knowledge becomes available (Ashe et al. 2006). Human imaging studies have also suggested the involvement of various parietal regions, including intraparietal sulcus, rostral inferior parietal, and precuneus (Grafton et al. 1998; Sakai et al. 1998; Bapi et al. 2006) in the generation and maintenance of internal plans. How the cerebral cortex integrates these plans with visual information to control movement via specific effectors remains a major unanswered question for understanding skilled action generation.

While the SRTT can reveal procedural learning (e.g., Grafton et al. 1998; Bapi et al. 2000; Poldrack et al. 2005), the finger-tapping movements involved may not provide rich enough data to uncover the effects of internal plans on voluntary movement, or to uncouple such effects from visual information. First, the SRTT measures only reaction time, a single component of naturalistic movements. Reaching movements allow the evaluation of many error features, such as amplitude, position, velocity and time. Multiple error features often arise from distinct cognitive channels (Bhat and Sanes 1998), and therefore may be subserved by partially separate neural substrates. By considering kinematics in multiple variables, we might better understand the effects of the internal plan on movements. Second, the SRTT shows significant performance effects from stimulus-response mappings or even perceptual probabilities as well as motor response probabilities (Lungu et al. 2004), suggesting that the SRTT is better suited to abstract or general sequence learning than to motor sequence learning per se.

The use of two-dimensional reaching movements introduces a number of issues not faced in the SRTT or other button-pressing tasks such as the 2 × 5 task (Hikosaka et al. 1995). Traditional experimental paradigms for point-to-point movements (e.g., Moran and Schwartz 1999; Diedrichsen et al. 2005) have many limitations such as small parameter spaces and strong interdependencies between variables (Paninski et al. 2004; Wang et al. 2007). A pursuit tracking task (Paninski et al. 2004) overcomes these challenges, but a stimulus with continuous random movement may be complex for human participants to memorize. In order to reduce learning demand while retaining continuous reach sequences sampled widely from the possible space of positions and direction changes, we developed a task having a continuously moving target stimulus that moves through a consecutive series of straight lines with fixed length and a traditional bell-shaped velocity profile, approximating the natural characteristics of linear point-to-point movements (Morasso 1981). These rigid control conditions reduce the sequential information to the direction of each element line, while still requiring naturalistic motor outputs.

One reason for the dearth of studies employing complex, continuous movements could relate to an absence of methods to distinguish trained actions from visually guided random actions in tracking tasks. Such methods would seem necessary toward characterizing the neural processes that differ between these modes of action. In this study, we evaluated kinematic differences and developed Bayesian and hidden Markov classification algorithms to discriminate between untrained (i.e., guided by sensory input) and trained (i.e., guided by both sensory input and an internal plan) movements, even when both modes occur as parts of a single continuous sequence.

When participants could use internal plan information in addition to sensory information, we expected to participants to eliminate or reduce errors caused by the visual feedback delay. Such performance improvement should appear in the form of reduced values for kinematic error measures such as position, velocity, and latency between hand cursor and target movement. However, we saw only subtle kinematic changes between the task's two control modes, the specifics of which varied across participants. The lack of consistent changes in kinematic error variables suggests that visual and internal plan information combine for reaching movement control upstream of the main source of common noise. Despite the participant-specific and often subtle extent of these changes, our hidden Markov classification algorithm could nonetheless identify the presence of an internal plan promptly and accurately.

Methods

Participants and apparatus

Fourteen healthy young adults (six male, age 23 ± 5) participated in the study. All participants were right-handed by self-report and had normal vision or corrected for normal. They gave informed consent to the procedures as approved by the Institutional Review Board of Brown University. Five of the 14 participants returned 5 ± 2 months after their initial session for a follow-up set of procedures.

Participants sat in a chair positioned in front of a computer controlled LCD panel (19″, 1,280 × 1,024 pixels) and an Intuos2 digitizing tablet (28.5 × 28.5 cm, 25.65° visual angle; Wacom Technology, Vancouver, WA, USA). A moving target (5 mm diameter, 0.45° visual angle) appeared on the LCD panel, and participants used the tablet's cable-free electronic stylus to control a position cursor (3 mm, 0.27° of visual angle). Each point on the tablet mapped to one point on the computer screen, such that distances were equal on the two devices. Participants held the stylus in a power grip throughout the experiment and moved the stylus only by shoulder and elbow movements, with the hand and elbow remaining in the horizontal plane atop the tablet. We determined the relative positions of the tablet and the participant's shoulder by allowing the participant to find a comfortable position with their arm in the above position, their hand resting on the center of the tablet. Custom software for Matlab 7.2 (The MathWorks Inc., Natick, MA, USA) controlled stimulus presentation and data collection; it updated target position at 62.5 Hz, and recorded *X*, *Y* location of the response cursor at 125 Hz.

Stimulus generation

We generated target trajectories using custom software written using Matlab. A trial's trajectory comprised a series of target cursor positions that moved the target through a series of straight-line movements. Each individual movement (segment) lasted 965 ms and covered 4.67 cm (4.2° visual angle), with an approximately bell-shaped velocity profile (Fig. 1b); see below for detailed discussion of trajectory creation. Each segment's mean tangential velocity was 5.02 cm/s, peak 11.35 cm/s. Within each segment, the highest target cursor velocity occurred at the segment midpoint. The number of segments per sequence varied between task conditions; the single "Repeat" sequence had six segments, the 50 "Novel" sequences each had six segments, and the 50 "Test" sequences had 18 segments each (see Task).

To generate a target trajectory for a candidate Repeat sequence, we chose a starting point randomly from a uniform distribution of *X* and *Y* coordinates within the workspace. We chose an endpoint for that segment by randomly selecting a movement direction from a uniform distribution of directions that would result in endpoints within the workspace, repeating to produce all six segments. We discarded a candidate sequence if it met any of the following rejection criterion: (1) it showed a significant correlation between *X* and *Y* position; (2) it showed a correlation between time and *X* position or *Y* position, i.e., its use of the workspace was unevenly distributed across time; (3) if the total distribution of positions showed a non-normal distribution around the center of the workspace in either *X* or *Y* space. If a candidate sequence passed this quantitative testing, we also examined it qualitatively, discarding it if it held any obvious memorable features, such as squares or other readily identifiable geometrical shapes.

We generated additional six-segment candidates as above for Novel trials, and tested them quantitatively as described above. In order to create 18-segment (17.37 s) sequences for Test trials, we spliced 2–10 (mean 6) extra segments onto the beginning of the Repeat sequence and 2–10 segments onto its end; we created each set of extra segments using the methods above. We subjected the full Test sequences to quantitative tests (1) and (2) only.

Task

Before the experiment, participants received instructions as to the overall task, as well as the order and content of the experimental conditions. Throughout the experiment, a participant's goal was to keep the position cursor as close as possible to the target cursor. A trial comprised a series of straight-line movements of the target cursor as described above with corresponding pursuit tracking. Inter-trial intervals lasted a minimum of 2 s, ending when the participant held the position cursor over the new target cursor for 1 s. Twenty or 50 consecutive trials of the same type comprised a block, depending on the task condition as described below. Intervals between blocks lasted at least 6 s, but participants were encouraged to rest before starting the first trial in the subsequent block. Trials were never overtly numbered or labeled.

A session comprised five task conditions. The first, "Novel," presented a different trajectory on each of 50 trials, to establish baseline performance for the continuous reaching task. The second, "Novel-Generation," established chance performance levels for generating trajectories from memory, by requiring the participant to recreate a trajectory seen only once previously. The third, "Repeat," allowed the participant to train on a single movement sequence by presenting it on each of 100 trials. The fourth, "Repeat-Generation," tested explicit learning by requiring

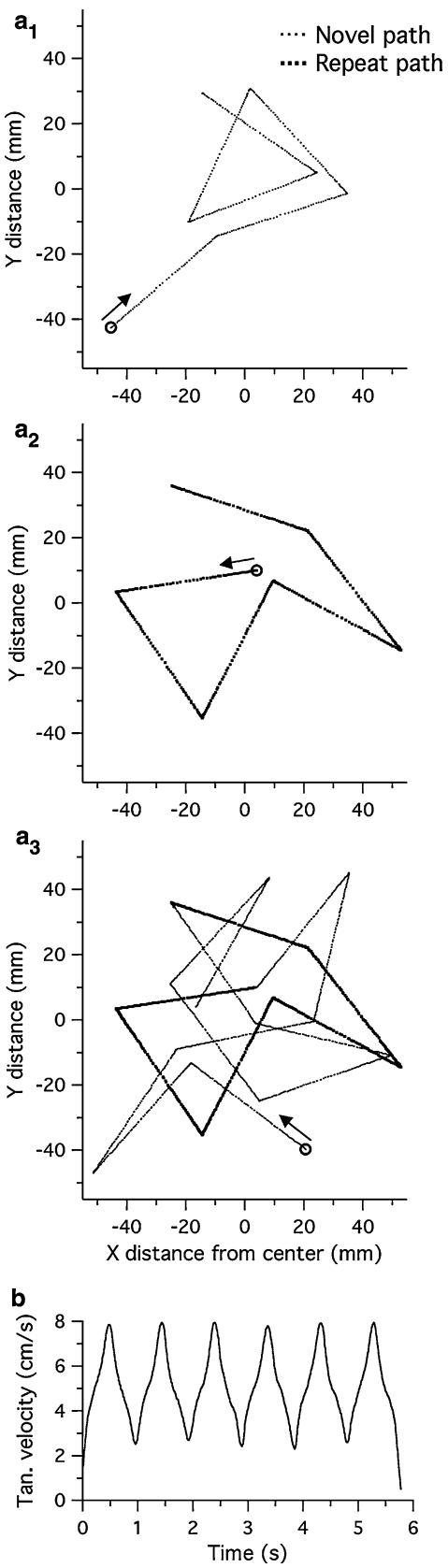

Fig. 1 Experiment design. For illustrations of trial types (**a1**–**3**), ▶ target moved continuously along path from start (*circle*) in motion direction (*arrow*), one point per display frame (16 ms). Although the complete paths are shown here, participants saw the target cursor in one point at a time. All distances measured from workspace center. Novel (**a1**) and Repeat (**a2**) trials each contained six segments. Test trials (**a3**) contained the Repeat path (*large dots*, six segments) flanked by Novel segments (*small dots*, 12 segments; here, six before Repeat and six after). Each segment has tangential velocity profile drawn from a normal distribution (**b**)

the participant to recreate the Repeat trajectory from the Repeat condition from memory. The final condition, "Test," required the participant to switch between pursuit of novel target cursor movements and familiar target cursor movements during a single long sequence.

The "Novel" condition (Fig. 1a$_1$) comprised a single block of 50 unique trials. Each trial comprised six segments (5.79 s duration). At the end of each trial, participants received error feedback in the form of the numerical value of the trial's root mean squared (RMS) position error in cm.

The second, "Novel-Generation" condition comprised a single block of 20 identical trials. The target cursor followed the last trajectory that had been shown during the Novel block; that trajectory only differed from other Novel trials by its placement at the end of the sequence block. During Novel-Generation trials, the target cursor became invisible after the first 0.5 s, forcing the participant to generate the trajectory from memory; the target cursor became visible again for the last 0.5 s. Participants received no error feedback at the end of each trial.

The third, "Repeat," condition (Fig. 1a$_2$) comprised 100 repetitions of the same trajectory, performed in two consecutive blocks of 50 trials each. This trajectory contained six segments (5.79 s duration). All participants received the same movement sequence for the Repeat condition, since preliminary data (not shown) showed no significant within-participant performance difference for different trajectories. At the end of each trial, participants received error feedback in the form of the numerical value of the trial's RMS position error, in cm. The fourth, "Repeat-Generation" condition comprised a block of 20 identical trials. As in the Novel-Generation, the target cursor became invisible for most of the trial; however, during Repeat-Generation, the invisible target cursor always followed the same trajectory as during the Repeat condition. Participants received no error feedback at the end of each trial.

The fifth, "Test" condition, comprised a block of 50 trials, each having 18 segments (17.37 s). Each trial comprised a random number (2–10) of novel segments, followed by the entire Repeat trajectory, followed by

another 2–10 novel segments (Fig. 1a$_3$). This divided each Test trial into three epochs: the first epoch of Novel movements (Novel-1), the Repeat epoch, and the second epoch of Novel movements (Novel-2). There were no overt differences between the epochs, but participants knew that each trial would contain the Repeat sequence. Participants pressed a key with their left hand when they became aware that they had entered or exited the Repeat sequence; i.e., when they identified a change in epoch. After each trial, participants assigned a 1–10 subjective accuracy rating to their key presses for the trial. Participants received no error feedback at the end of each trial.

Condition order was not randomized; in order to optimize performance at the complex task of the Test sequences, the Test condition immediately followed the Repeat conditions.

Five participants returned for a second session in order to present a version of the Test condition including catch trials, in order to evaluate success of classification models (see below) in the absence of any Repeat sequence. On the second session, participants received an abbreviated version of the above protocol, containing only two conditions. In a Repeat condition (Second-Repeat), they performed 50 trials of the same Repeat trajectory they had performed on their first session. In a Test condition (Second-Test), they performed 50 trials, of which 20% (ten trials) were Catch trials while the rest followed the Test trial format above without repeating any individual trials from the previous session. Catch trials were the same length as Test trials, but did not contain the Repeat sequence. Participants heard at session start that Test trials may or may not contain the Repeat sequence.

Data analysis

Of the 50 Test trials performed by each participant, we included 37.3 ± 9.0 (median 39.5) for further analysis. Exclusion criteria, indicating participant inattention or failure to perform the presented trial, included either of the following: (1) the participant failed to press the key (marking a change in epoch) twice; (2) if the participant's self-reported accuracy rating for the trial was more than two standard deviations below their own mean value; i.e., if the trial showed unusually poor performance or attention. These criteria allowed us to focus on trials that included conscious identification of the Repeat sequence.

We calculated *X* position error, *Y* position error, *X* velocity error, *Y* velocity error, and RMS position error between the position cursor and target cursor for each sample point (125 Hz). We calculated instantaneous velocity error by subtracting the response cursor velocity from target cursor velocity on each axis separately. For classification models, we smoothed these measures by averaging from a time window of 152 ms (19 sample points), centered on the current sample point. We determined lag (time error) values by identifying the peak of the cross-correlogram between target cursor velocity and position cursor velocity (904 ms width, 113 sample points) calculated at (i.e., centered on) each sample point, after the previously described smoothing.

Because participants knew the length of the Repeat sequence, and therefore could predict the transition from the Repeat epoch to the Novel-2, all Test trial analyses compared the Novel-1 epoch against the Repeat epoch. The task's attentional requirements differed widely between Novel-1 (in which the participant needed to identify the beginning of the Repeat sequence) and Novel-2 (in which the participant had no such need). As such, we decided not to group Novel-1 and Novel-2 data. Data and analyses for the transition between the Repeat epoch and Novel-2 are not shown, and all "Novel epochs" refer to Novel-1, unless otherwise specified. Furthermore, we removed the first segment (0.965 s) of Novel-1 from analyses to avoid effects specific to the start of movement.

We generated latency times by adding the full window length used for data averaging to the time elapsed between task epoch change and estimated epoch change. For Bayesian classification (see below), we based the epoch change time on a measure of when the model identified the epoch reliably. For this purpose, we defined a reliable classification as the start of a period of consistent classification of a calculated minimum length. We determined that minimum length as the smallest window that produced less than 5% misclassifications (false positives) during the Novel epoch for that participant.

We did not assess key-press data in detail due to unreliability of precise timing in Matlab time-stamp recording. During data collection, the Matlab software introduced noise while determining the time of discrete events (e.g., key-presses and task epoch change times). We knew the ground truth for task epoch change times, so we determined the Matlab-introduced error by comparing those values with the online-measured task epoch change times. We corrected population measures of key-press times by the mean Matlab-introduced error for each participant (−17 ± 99 ms across participants).

All significance values represent results from *t* tests unless otherwise specified.

Classification algorithms: Bayesian

Bayesian maximum-likelihood classification allowed estimation of a movement's state (control mode; trained or untrained), based on the probability of data distributions occurring under each state and the probability of each state occurring during training. We built a model of Repeat and

Novel movements and tested this model by attempting to make instantaneous identifications of the control mode during a participant's movements. We built a Gaussian mixture model using X and Y position error, X and Y velocity error, and time lag, and estimated its parameter $\Theta = \{P_S,\ \mu_S,\ \Sigma_S$: S = repeat or novel$\}$ from all trials except the one currently being tested (cross-validation). See Press (1989) for a detailed discussion of Bayesian classification theory and implementation.

In order to estimate the model parameter Θ, we chose the one that maximizes the joint probability [maximal likelihood estimate (MLE)]. The MLE is the empirical frequency for each state P_S, and empirical mean μ and covariance matrix Σ for each Gaussian component. Given training data D,

$$\begin{aligned}\left(P_S^*, \mu^*, \Sigma^*\right) &= \arg\max_{P_S,\mu,\Sigma} P(D, S|P_S, \Sigma) \\ &= \arg\max_{P_S,\mu,\Sigma} P(D|S, \mu, \Sigma)P(S|P_S)\end{aligned}$$

where (S, D) are the known task states and data for all training trials, and * indicates an estimated value.

We inferred the state (Repeat vs. Novel condition) S^* using data D_{test}

$$\begin{aligned}S^* &= \arg\max_S P(S|D_{\text{test}}; \Theta) \\ &= \arg\max_S P(D_{\text{test}}|S; \Theta)P(S|\Theta)\end{aligned}$$

We implemented these equations by calculating $P(D_{\text{test}}|S)$ from the formula

$$P(D_{\text{test}}|S) = \frac{1}{(2\pi)^{\left(\frac{5}{2}\right)}|\Sigma_S|^{\left(\frac{1}{2}\right)}} \exp^{-\frac{1}{2}\left(D_{\text{test}}-\mu_S^*\right)^T\left(\Sigma_S^*\right)^{-1}\left(D_{\text{test}}-\mu_S^*\right)}$$

and drawing P_S from the task design.

Classification algorithms: hidden Markov models

Hidden Markov models (HMM) estimate the time of a state change, based on the distributions and past history of state-dependent variables (see, e.g., Rabiner 1989). We assumed that the internal states formed a Markov chain; that is, the current state depended only on the immediately previous state, and generation of the observed errors depended probabilistically only on the current state. The unobservable Markovian internal state sequence and the observable kinematic error sequence together form a HMM that provides a principled estimate for the occurrence of state changes.

We trained the HMM on all Test trials except the one being evaluated (cross-validation). We trained the model by estimating the transition probabilities for the Markov states and the dependency parameters for distributions of kinematic errors during each state. Because we knew the ground truth of task condition for training trials, we estimated transition probability from transition frequency over time, and estimated the Gaussian means and covariance matrix from the empirical means and covariance matrix for performance variables during each condition.

$S_{1:T} = (S_1, S_2, \ldots, S_T)$ denotes the sequence of internal states and $D_{1:T} = (D_1, D_2, \ldots, D_T)$ denotes the kinematic errors. The joint probability of

$$\begin{aligned}P(D_{1:T}, S_{1:T}|\Theta) &= \prod_{t=1}^{T} P(S_T|S_{T-1}, \Theta)P(D_t|S_t, \Theta) \\ &= \prod_{t=1}^{T} P(S_T|S_{T-1}, A)P(D_t|S_t, \mu, \Sigma)\end{aligned}$$

where $\Theta = \{A,\ \mu,\ \Sigma\}$ are the transition probability between the different states, and the Gaussian component's means and variances, which specify the dependency between the observed kinematic error and the internal state, respectively. We estimate the optimal Θ^* that maximizes the joint probability $P(D, S|\Theta)$ for the training data set.

After training, we decoded the optimal internal states for the test data that maximizes the posterior probability

$$\begin{aligned}S^* &= \arg\max_S P(S|D_{\text{test}}; \Theta) \\ &= \arg\max_S P(S, D_{\text{test}}; \Theta) \\ &= \arg\max_S \prod_{t=1}^{T} P(S_t|S_{t-1}; A)P(D_t|S_t, \mu, \Sigma)\end{aligned}$$

We achieved the decoding using the Viterbi algorithm, a particular case of the dynamic programming algorithm (Geman and Kochanek 2001). This algorithm uses the sequential structure of the HMM, enabling us to decode the optimal hidden states in linear time. In order to implement this decoding, we calculated the maximal log likelihood C as

$$\begin{aligned}C(t, S_t) = \max_{S_{t+1}} &\log P(S_{t+1}|S_t, A) + \log P(D_t|S_t, \mu, \Sigma) \\ &+ C(t+1, S_{t+1})\end{aligned}$$

for each time t and possible current state S, given data D_t. We calculated this and the following formula in a tail-to-head (reversed) direction for ease of computation, since the reversed chain is also Markovian. We also save the possible optimal state R at each instance t as

$$\begin{aligned}R(t, S_t) = \arg\max_{S_{t+1}} &\log P(S_{t+1}|S_t, A) + \log P(D_t|S_t, \mu, \Sigma) \\ &+ C(t+1, S_{t+1})\end{aligned}$$

We determined the final optimal state S^* by choosing the values of S (that is, the columns of R) from the value of R at the previous t, via

$$S_{t+1}^* = R\left(t, S_t^*\right).$$

Note that this returns state selection to a head-to-tail (forwards) direction.

Classification algorithms: discriminant analysis

To compare the above analyses against a simpler classification scheme, we also classified our raw data using a variant of linear discriminant analysis (LDA), which finds the linear combination of features separating classes of events. To do so, we wished to classify our data based solely on RMS position error, a traditional kinematic measurement. We used Fisher's linear discriminant (FLD) algorithm, a variant of LDA that functions even on data that are not normally distributed, as was the case for RMS position error here. We implemented FLD using a Matlab script released by Nuo Li ("LDAplane.m", http://www.mit.edu/~linuo/Matlab%20Scripts.html).

Results

Learning

We first evaluated whether participants learned movement sequences through position error reduction, using a position RMS error (RMSe) measure normalized within subjects to the first Novel trial. Figure 2a illustrates the success of participants to learn the task and the Repeat sequence. The large error reduction occurred in the first ten trials for the novel task, and did not appreciably decline further (Fig. 2a, thin trace) since the RMSe for the last 40 trials did not differ significantly from zero slope ($P > 0.99$). During the Repeat condition, mean error initially declined beyond that for the Novel task (Fig. 2a, thick trace), while the last 80 trials did not differ significantly from zero slope ($P > 0.99$).

Participants returning for a second session did not perform the Novel condition on that day, but still started their Repeat trials (Second-Repeat condition; Fig. 2a, dotted trace) at approximately the same error level as the Repeat trials on the first session; for participants who performed both sessions, we could not distinguish mean RMSe between the first Repeat trial and first Second-Repeat trial ($P = 0.187$). There was some evidence of retained learning of the specific sequence, in that participants performed significantly better during the second session after performance stabilization; trial mean RMSe values were significantly lower than for the original Repeat condition during its stable region ($P \leq 0.0001$).

We next compared performance among the last 18 trials of Novel-Generation, Repeat-Generation and Repeat conditions to demonstrate the presence of explicit learning (Fig. 2b). We examined only 18 trials because participants sometimes misunderstood the directions for the first 1–2 trials. Novel-Generation trials had a mean RMSe of 45.46 ± 28.65 mm, while Repeat-Generation had mean 12.68 ± 8.72 mm, and the Repeat trials had mean 7.99 ± 7.94 mm. Pairwise tests showed that each distribution differed significantly from each of the others (Kolmogorov–Smirnov (K-S) tests, $P \leq 0.0001$), confirming that these conditions yielded different amounts of kinematic error.

Performance

Figure 3 depicts performance during one sample Test trial. Figure 3a shows the paths of both cursors over time, and Fig. 3b shows velocity profile, while Fig. 3c–f illustrate time courses of individual error measures throughout the same trial. Shifts in performance between Novel and Repeat epochs are subtle at best; while some changes in mean or variance can be observed, ranges remain highly overlapped. In order to quantify these changes, we examined distributions of kinematic variables across Test trials within each participant.

Individual participants showed minimal but statistically significant performance changes across conditions. Figure 4 summarizes performance data across all Test trials for one participant, showing error measures calculated from all data points. All the distributions in Fig. 4 exhibited significant differences between Repeat and Novel epochs (K-S test,

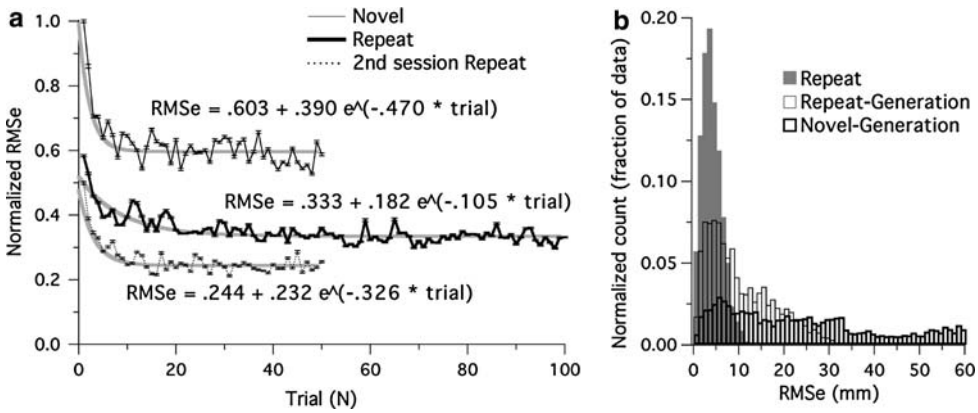

Fig. 2 Evidence for learning of movement sequences. **a** Learning curves, normalized to participant mean RMS position error on first Novel trial. *Gray lines* indicate fit equations. **b** RMS position error distribution for 18 Generation (explicit memory test) trials and normal (non-Generation) Repeat trials. Values normalized within subjects. Uppermost 1% of points in distributions not shown

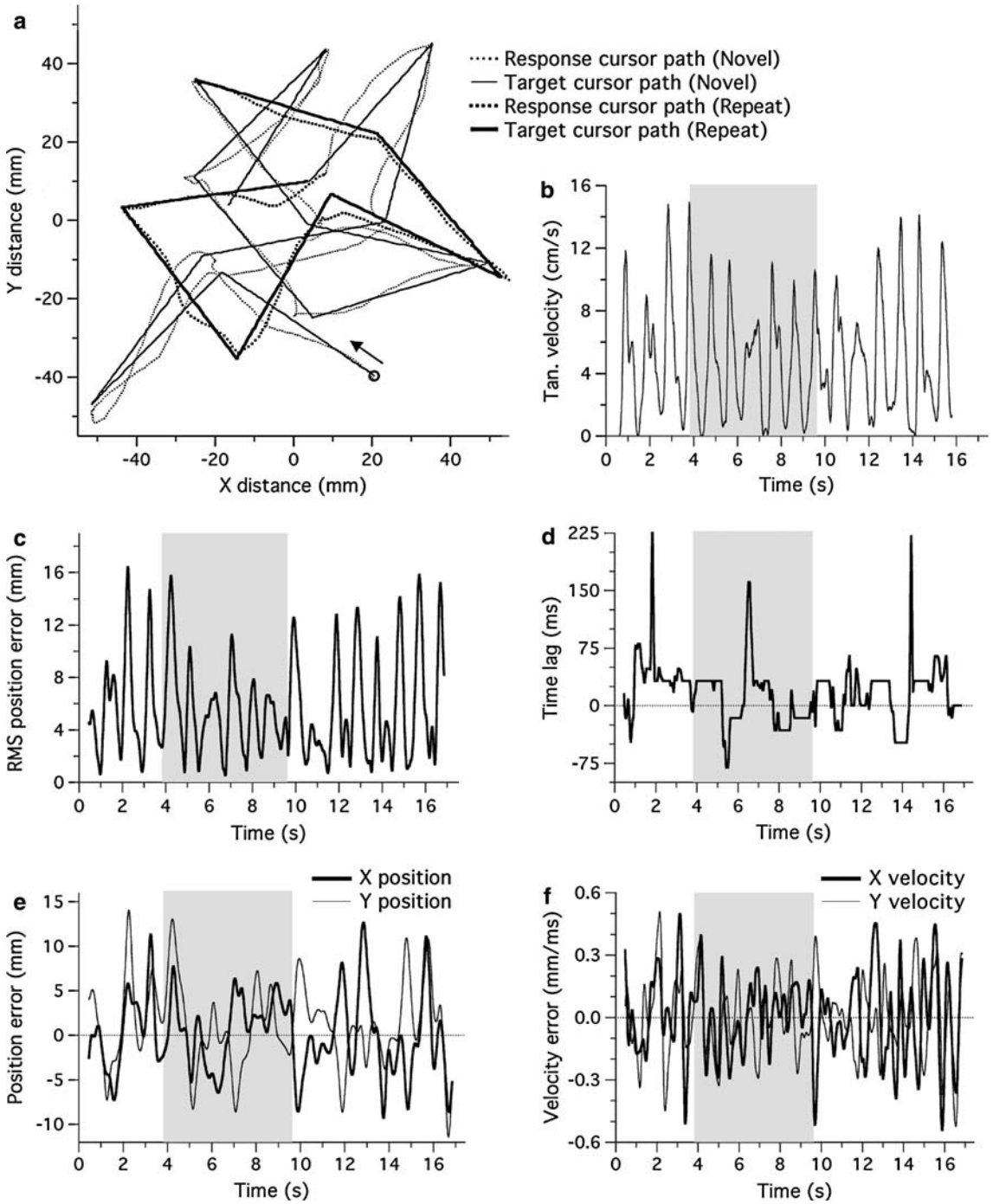

Fig. 3 Time course of one sample Test trial. *Shaded regions* indicate the Repeat epoch for variable displays (**b–f**). **a** Cursor paths; *arrow* indicates direction of motion. Distances measured from workspace center. **b** Tangential velocity. **c** RMS position error. **d** Time lag, computed from cross-correlogram peak. **e** *X*/*Y* position error. **f** *X*/*Y* velocity error

$P < 0.001$), even though they showed considerable overlap. We tested for differences between the two distributions by calculating the proportion of the area under either curve shared by both curves. Across variables, the areas under the curves for the two conditions overlapped 82.6 ± 5.5%. Across participants, these curve areas overlapped 81.9 ± 2.2%. We reduced the data by averaging error measures within individual movement segments, and examining the distributions of all movement segment averages (not shown; for one example participant, $n = 267$ for Novel, $n = 322$ for Repeat). Examining this segment-mean performance, error measure curves overlapped by 74.6 ± 2.1% across participants. Using segment means, each individual had significantly different distributions

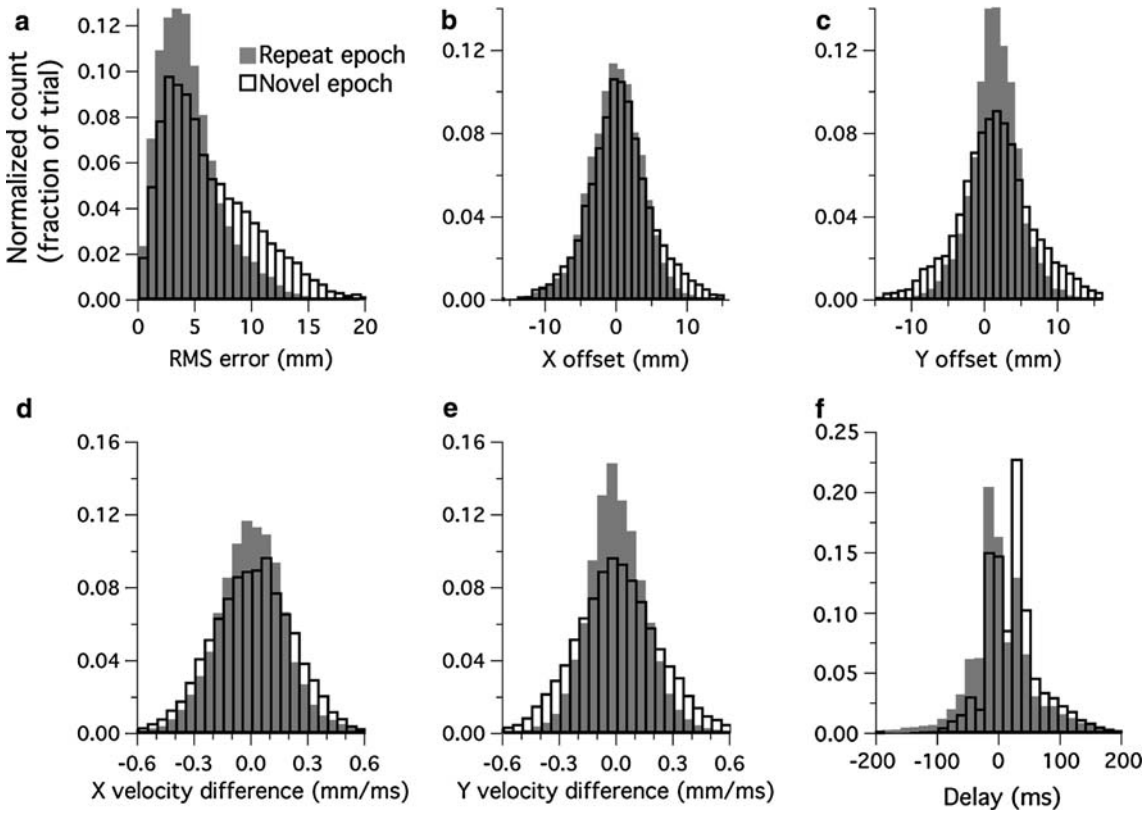

Fig. 4 Distribution of error measures across Test trials for one representative participant, showing all data points ($n = 26{,}464$ for Novel, $n = 33{,}131$ for Repeat). **a** RMS position error (80.4% overlap between curves). **b** *X* position error (92.2% overlap). **c** *Y* position error (78.7% overlap). **d** *X* velocity error (86.1% overlap). **e** *Y* velocity error (79.4% overlap). **f** Time lag (78.6% overlap)

between epochs for 3.07 ± 0.83 of the five variables (K-S test, $P \leq 0.05$); *X* position error, *Y* position error, *X* velocity error, *Y* velocity error, and time lag. Each of those variables differed significantly for 8.6 ± 5.7 of the 14 participants (median 11). We did not examine RMSe directly since RMSe could not be used for many classification algorithms due to its non-normal distribution, but see the following paragraph for analysis of unsigned errors. Overall, each participant showed a different pattern of differences in the measured variables between epochs; MANOVA revealed a significant effect of participant on each variable ($P \leq 0.05$ for *Y* velocity; $P \leq 0.001$ for other variables) and a significant interaction effect between epoch and participant for each variable ($P \leq 0.001$). Taken together, these data suggest that distributions of error measures changed between epochs of our Test task, but only to a small extent. When two distributions show 79–89% (raw data) or 66–84% (segment means) overlap between epochs, error measures within a single data point or even single segment cannot suffice to distinguish between epochs.

We also examined unsigned position and velocity errors, to determine whether the use of signed errors concealed any patterns in absolute magnitude values. When using unsigned errors, each individual participant had error measure distributions that overlapped 83.4 ± 2.4% between epochs, or 73.7 ± 2.6% using segment-mean data. These values did not differ significantly from the data above ($P = 0.460$ for raw data, $P = 0.789$). For unsigned segment-mean data, each individual had significantly different distributions between epochs for 4.6 ± 0.8 of the five variables, significantly more than for signed data ($P < 0.001$). Each variable differed significantly for 13 ± 0.8 of the 14 participants, not significantly different than for signed data ($P = 0.131$). As with RMSe, we did not use unsigned position and velocity errors for Bayesian or HMM classification algorithms due to their non-normal distributions.

When the condition changed between Novel and Repeat trials, each participant still showed significant overlaps between distributions of error measures. Error measure curves overlapped by 68.4 ± 3.9% across participants, significantly less than for Test epochs ($P < 0.005$). Examining data by segment-means, distributions overlapped by 54.6 ± 2.9%, not significantly less than for Test epochs ($P = 0.095$). Importantly, RMS position error overlapped significantly and drastically less between trial types than any other variable; all other variables overlapped 63.6 ± 3.0%, while RMS position error overlapped only 9.4 ± 11.7% ($P < 0.001$).

Classification

We used various classification algorithms to demonstrate that the error variables differed sufficiently between conditions to allow us to distinguish the two conditions. In order to identify the condition at any specific time, we classified sliding windows averaged over 152 ms of data, as summarized in the receiver operating characteristic (ROC) plot of Fig. 5. All accuracy values represent the proportion of time windows wherein the model correctly classified the data across participants (group mean ± SEM). In addition to classification success rates across whole trials, we also report Novel-epoch accuracy in terms of hit rates (Novel data classified correctly), and Repeat-epoch accuracy in terms of false positive rates (Repeat data classified incorrectly).

We first performed FLD, a method designed to classify data based on the traditional measure of RMSe (Fig. 5, black circle). FLD produced an overall accuracy rate of 60.1 ± 4.5% ($P \leq 0.001$). However, looking at each epoch separately, FLD had a hit rate of 64.5 ± 6.6% ($P \leq 0.001$), but a false positive rate of 45.6 ± 10.4%, not significantly different from chance ($P = 0.134$). We further confirmed the inability of FLD to identify the control mode during the task at hand by its chance performance during Catch trials (see below).

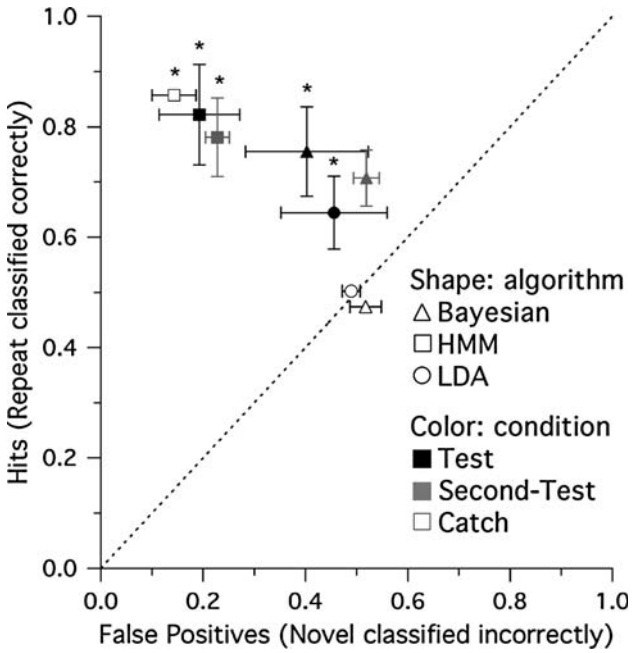

Fig. 5 ROC plot for algorithm classification rates for various data sets. Second-Test data represents Test trials from session with intermixed Catch. For Catch trials, the single available accuracy value determines position on both axes. All values group mean ± SEM. *Dashed line* indicates complete chance performance. *Asterisk* indicates an algorithm that performed better than chance on all axes

In order to maximize the ability to draw classification information from the data, we switched to parametric models that undergo explicit training; we started with Bayesian classification. We used all five of the kinematic error variables above (*X* position error, *Y* position error, *X* velocity error, *Y* velocity error, and time lag) to train these algorithms; as described below, each variable provided significant information in at least one setting. Under these conditions, Bayesian classification for epochs of Test trials (Fig. 5, black triangle) had a hit rate of 75.5 ± 8.1%, significantly higher than chance ($P \leq 0.001$). Bayesian classification had a false positive rate of 40.3 ± 12.0%, significantly lower than chance ($P \leq 0.01$). Overall, Bayesian classification correctly identified the epoch for 68.7 ± 4.4% of the Test data ($P \leq 0.001$); participant means ranged 61.8–75.7% correct, all significantly better than chance ($P \leq 0.001$).

We also applied HMM to classify Test trial data (Fig. 5, black square), which look at the history of the data to identify the time of a state change, rather than classifying every data window independently. HMM yielded a hit rate of 82.2 ± 9.1% ($P \leq 0.001$) and a false positive rate of 19.3 ± 7.9% ($P \leq 0.001$); its overall Test accuracy reached 81.7 ± 5.1% ($P \leq 0.001$), significantly higher than the accuracy rate for Bayesian classifications ($P \leq 0.001$). Participant means ranged 70.2–89.8% correct, all significantly better than chance ($P \leq 0.001$).

Some participants returned to perform a version of the Test task (Second-Test) that included 20% catch trials that did not contain the Repeat trajectory. We trained classification algorithms on the Test trials and tested on both Test and Catch trials. Only HMM successfully classified all data more often than chance during these conditions. Bayesian classification failed to reliably classify data correctly during Test trials (Fig. 5, gray triangle). While hit rates during Repeat epochs were better than chance (70.7 ± 5.1%, $P \leq 0.001$), as were overall accuracy rates (60.6 ± 3.9%, range 55.8–64.3%, $P \leq 0.01$), false positive rates were no different from chance (51.9 ± 2.5%, $P = 0.15$). For the Catch trials themselves, the lack of a Repeat epoch meant only a single accuracy measure (correct rejections vs. false positives) was available. Bayesian classification rates for the Catch trials did not differ significantly from chance (white triangle, 47.9 ± 2.9%, range 46.3–54.4%, $P = 0.66$). We confirmed the inadequacy of FLD for this task by showing that it could not classify Catch trials better than chance, whether we performed FLD on either RMS position error (white circle, 50.6 ± 1.7%, $P = 0.453$) or the same set of kinematic variables as other methods (not shown, 52.7 ± 5.4%, $P = 0.325$). HMM testing, however, correctly classified data during both Test trials (gray square; overall accuracy 77.7 ± 3.3%, range 72.3–80.6% $P \leq 0.001$) and Catch trials (white square; 85.7 ± 4.3%,

range 75.6–93.2%, $P \leq 0.001$), demonstrating its superiority for classifying data of this type. Overall, classification results suggest that performance differences exist between the different conditions, but performance remains quite similar between them.

Contribution of variables

We performed a stepwise regression to measure how much each possible factor contributed to the classifications made across participants using the Bayesian and HMM algorithms. We considered as possible factors: *X* position error, *Y* position error, *X* velocity error, *Y* velocity error, time lag and participant. Considering both Test epochs together (i.e., how each factor contributed to classification accuracy across entire Test trials), *X* velocity error failed to contribute significantly to either algorithm (Bayesian $P = 0.271$, HMM $P = 0.126$, other variables $P \leq 0.001$), and *Y* velocity error failed to contribute significantly to HMM classification ($P = 0.065$). However, when looking at Repeat and Novel epochs separately, all six factors contributed significantly ($P \leq 0.001$) to Repeat epochs for both algorithms, and only *Y* velocity error failed to contribute significantly to Novel epochs (Bayesian $P = 0.827$; HMM $P = 0.116$). Each kinematic variable therefore did contribute to the models under some circumstances, although velocity error variables contributed less often. As discussed below, the change in relative importance of various error measures can be described as a change in motor strategy between the two epochs, as the motor system optimizes different movement parameters.

We ranked the factors in order from highest to lowest confidence of a contribution (*P* value), in order to observe how the relative contributions of each factor differed between algorithms for the group. For the Bayesian classification, the factors listed from highest to lowest contribution, considering both epochs together, were: *X* position error, *Y* velocity error, participant, time lag, *Y* position error, *X* velocity error. For the HMM classification, the factors listed from highest contribution to lowest were: participant, *Y* position error, time lag, *X* position error, *Y* velocity error, *X* velocity error. This showed that the HMM was more sensitive to between-participant differences. More broadly speaking, the two algorithms may draw different amounts of information from diverse variables because they differ not only in formulae, but in the HMM's characteristic ability to look at the history of data, and hence be more sensitive to changes in a particular variable over time.

In order to more clearly demonstrate the contribution of each variable to predict the movement type, we performed Bayesian and HMM classifications for each participant using only one of the five above variables at a time. These single-variable classifications usually performed better than chance (61/70 Bayesian and 69/70 HMM, multiple-comparison corrected $\alpha = 0.0007$ each). Averaged across participants, single-variable Bayesian classifications performed with group mean ± SEM accuracy 57.8 ± 3%, significantly above chance ($P \leq 0.001$); HMM classifications performed with 74.3 ± 4.3% accuracy, also above chance ($P \leq 0.001$). However, these relatively high values resulted in part from individual variables classifying well during only one of the two Test epochs; looking only at Novel epochs, far fewer single-variable classifications performed better than chance, especially for Bayesian classification (45/70 Bayesian and 66/70 HMM, $\alpha = 0.0007$). Furthermore, these single-variable classifications performed consistently worse than classifications based on the full set of five kinematic variables described above, both individually (70/70 Bayesian and 66/70 HMM, $\alpha = 0.0007$) and averaged across participants ($P \leq 0.001$ for both Bayesian and HMM; see below for five-variable classification rates). Figure 6 shows the results of single-variable classification for one representative participant, using each kinematic variable during each Test epoch, demonstrating the ability of many variables to classify well during only a single epoch. The same variables favored different epochs for different participants; even for the 5/10 variable-algorithm pairs with significant across-participant biases, the variable went against its across-participant trend for an average of 3.0 ± 0.7 participants (21 ± 5%). For both Bayesian and HMM classifications, ANOVA revealed a significant epoch-by-variable interaction effect ($P \leq 0.001$) for each participant, showing that participants favored different kinematic variables during the two epochs, and that the variable selected was not consistent across participants.

Algorithm variations

The way that different variables contributed differently for different task epochs led to the concern that the true performance errors may be best found in a reference frame other than Cartesian (i.e., *X* and *Y*) axes. For example, if participants primarily made errors in the length of their movement segments, this would appear as a different Cartesian error depending on the segment direction. In order to address this concern, we decomposed error values into a movement-centered reference frame, such that our two axes were along and perpendicular to the movement axis. These data resulted in accuracy rates, across whole test trials, of 61.3 ± 4.6% for Bayesian classification and 74.4 ± 5.3% for HMM, both significantly lower than the accuracy rates described above for the Cartesian data ($P \leq 0.001$ for each algorithm). These consistently and significantly worse classification results suggest that

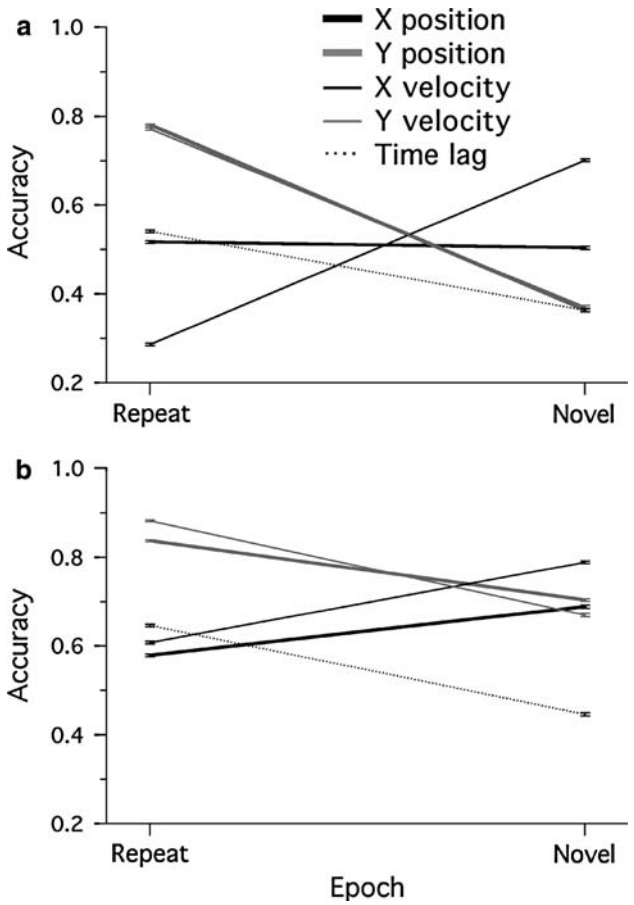

Fig. 6 Classification accuracy (group mean ± SEM) from algorithms trained on individual variables, across trials for one representative participant. Accuracy rates represent the proportion of time windows in which the algorithm correctly classified the data. **a** Bayesian classification. **b** Hidden Markov model classification

Cartesian coordinates better capture participants' performance errors than movement-centered coordinates. We describe our reference frame as "Cartesian," but this label only means that we captured movement errors in the form of vertical and horizontal errors with respect to the movement plane. While we did not measure eye movements, we can reasonably assume that participant eye movements stayed on the visual workspace for analyzed trials, eliminating any practical difference between the target-movement and visual planes. Therefore, our coordinate system is consistent with models that compute hand control in a vision-centered coordinate frame with its origin at the shoulder (e.g., Shadmehr and Wise 2005).

As an additional concern, our classification models did not initially take advantage of the non-stationarity in the data; as Fig. 3 shows, position and velocity errors decreased at the beginning and end of each segment (when target speed was low), compared to the middle (when target speed was high). In order to integrate this non-stationarity into our models, we added target velocity to the set of variables used for Bayesian and HMM classification. Because target velocity did not change between Novel segments and Repeat segments, this would only allow our models to take advantage of any correlation between target velocity and kinematic errors. Classification accuracy rates from these models did not differ significantly from our original models. Across whole Test trials, these rates were 68.8 ± 4.5% for Bayesian classification ($P = 0.936$ from original classification data above) and 81.9 ± 5.0% ($P = 0.890$) for HMM. Decomposing target velocity into separate X and Y velocity variables produced similar results, with classification rates of 69.6 ± 4.2% ($P = 0.578$) for Bayesian and 81.6 ± 6.0% ($P = 0.971$) for HMM. These results suggest that any non-stationarities in the kinematic data could not meaningfully contribute to these classification methods.

Latency to classification

In addition to estimating the capability of classification schemes to correctly identify the class of movement, that is, Novel versus Repeat, we also assessed how soon our primary classification methods could reliably identify the switch between classes. Note that the term "latency" here represents the time between task epoch change and first algorithm classification of the Repeat epoch, as distinct from "lag," the time error between target and response cursor movements.

Bayesian classification resulted in mean latency from the onset of the embedded sequence to first correct classification of 840 ± 746 ms across participants, while the HMM method resulted in a latency of 244 ± 696 ms. The mean latency of participant key press responses was 1,099 ± 611 ms. The HMM method produced latency values significantly lower than either Bayesian latencies or participant key presses, but Bayesian and participant key press latencies did not differ significantly ($P \leq 0.05$, $P \leq 0.01$, and $P = 0.325$, respectively). Figure 7a shows HMM and response latency by participant, demonstrating that, for most participants, HMM latencies remained consistently lower than key press latencies, even taking into account an approximate key press reaction time of 150 ms (e.g., Nissen and Bullemer 1987). Figure 7b shows the distribution of latency values within one sample participant, with many values around the means but also many outliers. Latency times from sessions including catch trials showed the same pattern; Bayesian classification resulted in 1,465 ± 507 ms latency, HMM 278 ± 502 ms, and key press latencies 2,499 ± 1,836 ms (pairwise comparisons ordered as above; $P \leq 0.05$, $P \leq 0.05$, and $P = 0.259$). Note that negative latency values could occur, if an algorithm incorrectly identified the onset of the embedded sequence before its actual start time.

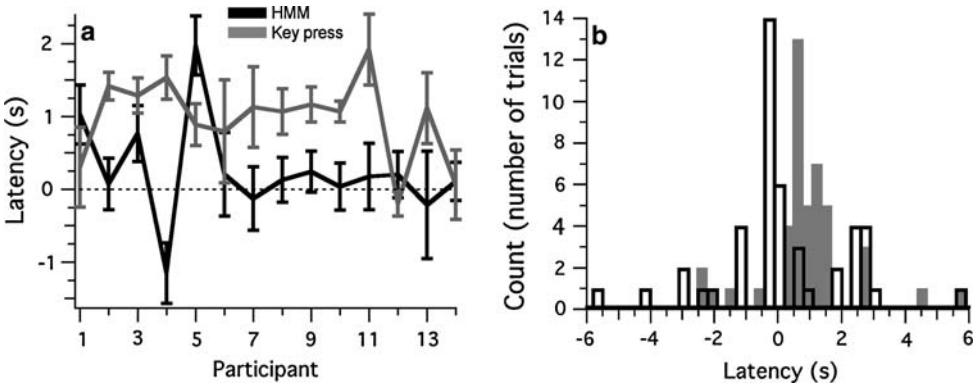

Fig. 7 Distribution of latencies for HMM and participant key press responses. **a** Across participants (group means ± SEM). Displayed key press response latencies have been reduced by 150 ms from original data, to account for reaction times. **b** Histogram of trial latencies for one sample participant. Two highest latencies for key presses (9,511, 7,476) not shown

We performed additional tests to determine whether negative latency values (i.e., false positive classifications) occurred when Novel segments had similar orientations or positions to the beginning of the Repeat sequence. We defined a "similar orientation" as a segment orientation within 30° of the first Repeat segment's orientation. We defined a "similar position" as a position with a distance from the first Repeat point less than one standard deviation below the trial mean distance. Similar orientations occurred on 11.7 ± 11.4% of false positives for Bayesian classification, and on 4.4 ± 3.8% for HMM, not significantly different from a chance rate of 8.3% ($P = 0.790$ for Bayesian, $P = 0.585$ for HMM). These results for the two algorithms were significantly but marginally different ($P = 0.059$). Similar positions occurred on 8.6 ± 7.1% of false positives for Bayesian classification, and on 16.3 ± 8.4% for HMM, not significantly different than a chance rate of 15.8% ($P = 0.062$ for Bayesian, $P = 0.893$ for HMM). These results differed significantly between the two algorithms ($P = 0.135$). Therefore, false positive classifications were not significantly associated with any simple overt similarities between Novel and Repeat segments,

The HMM mean latency of 244 ± 696 ms is lower than the duration of a single movement segment, unlike keypress latency. This might give rise to the concern that the model seems to recognize the sequence impossibly quickly, while keypress latency provides a more realistic estimate of identification time. While participants can covertly express procedural learning before they can overtly declare it (e.g., Nissen and Bullemer 1987), we nonetheless wish to address this potential concern. If the first Repeat segment contained a non-trivial length of time in which the algorithm has incorrectly identified a state change, the first segment would then differ from the later segments in some fashion. To verify that the low latency times did not result from anything specific and unusual about the first segment of the Repeat epoch, we shuffled the order of the segments within the Repeat epoch before classification. Because the HMM takes account of data history, it distinguishes between momentarily identical data at different times in a sequence. These shuffled data had a latency of 439 ± 665 ms, not significantly different from the 244 ± 696 HMM latency reported above ($P = 0.454$). We also tried removing the first Repeat segment before classification. These data, which omitted any issues arising from the first segment, had a latency of 605 ± 627 ms, also not significantly different from the original HMM latency reported above ($P = 0.161$). Therefore, the HMM identified a behavioral change equally rapidly whether or not that behavior could include a transitional period, in which behavior demonstrated Novel characteristics despite the task's Repeat epoch. This suggests that any such transitional period must be minimal.

We next examined the possibility that algorithms, or the behavior they analyze, could predict the onset of the Trained epoch from the passage of time during a trial. In the most extreme example, since the epoch change could occur no later than the start of the tenth segment, algorithms or participants could potentially predict the exact epoch change time if the Novel-1 condition continued after nine segments. A negative correlation between latency time and epoch change time would suggest that the algorithms or behavior predicted epoch change based on time elapsed within each trial. We found no significant correlation for latency times calculated from Bayesian classification ($r = -0.16$, $P = 0.732$). We found a significant positive correlation for HMM latency times ($r = 0.75$, $P \leq 0.05$), representing less accurate predictions when the epoch changes occurred late in the trial, perhaps due to participant attention waning or error accumulation across a trial.

Discussion

Based on the current results, we have described computational approaches allowing differentiation of performance of visually guided actions from those using vision and internal guidance. Simple analyses of performance cannot

easily differentiate these two conditions, even though the neural substrates presumably vary greatly between conditions with and without prior knowledge. For complex sequences of reaching movements, novel and familiar sequences with similar properties were difficult to differentiate because of the high level of independent kinematic error (motor noise) present in both movement types regardless of the presence or absence of an internal plan, as well as strong similarity in the way they were performed. While participants did change their motor strategy between Novel (untrained) and Repeat (trained) epochs, the distributions of performance variables overlapped highly for all variables in all participants. These and other subtle kinematic differences allowed HMM to successfully classify the control mode with $81.7 \pm 5.1\%$ accuracy and within 244 ± 696 ms. The slight kinematic changes between epochs suggest that most of the motor error for reaching movements arises from later, common stages in motor control, downstream from where visual and internal plan information combine.

Internal plan form

An internal plan for movement control may take the form of a procedural model of muscle activations, compared with proprioceptive feedback (e.g., Ariff et al. 2002) and efference copy information (e.g., Duhamel et al. 1992); it may act as a memorized sequence of explicit cues, providing cognitive information about upcoming targets before visual input becomes available; or it may include both types. The decrease in position error from Novel-Generation to Repeat-Generation revealed the effect of implicit and explicit learning combined, with participant verbal reports entailing at least some role of explicit knowledge. The increase in error from Repeat to Repeat-Generation showed unsurprisingly that some of the training-based improvement depended on visual information, and that vision reduced differences between trained and untrained movements. More interestingly, the performance differences between the Repeat and Repeat-Generation tasks suggested a significant cognitive component to the form of the internal plan. If the internal plan had been entirely procedural (implicit), the lack of error feedback would not have impeded the participants' ability to move using that plan. Alternatively, this effect could arise from partial implicit learning that requires complementary visual information. However, numerous studies (e.g., Karni et al. 1995; Grafton et al. 1998) have demonstrated that explicit learning occurs before procedural learning in non-covert sequence learning tasks. Force field adaptation studies have long established procedural information as a major component of reach movement control (e.g., Wang et al. 2001). The combination of procedural and cognitive components in our participants' internal plans suggests that their plans incorporated information about both cues and movements.

It is worth remembering that our main task did not strictly discriminate between visual and internal plan information. Participants could use visual information during either of the main task modes (Repeat or Novel); we did not attempt to isolate the role of visual information, except for brief analysis of the Generation trials discussed above. Instead, in order to evaluate the advantage of having an internal plan to generate visually guided actions and action sequences, we isolated its contribution by switching between Repeat and Novel epochs, which only differed in terms of the presence or absence of such prior knowledge.

Error measures

For complex movement sequences such as those evaluated here, learning primarily produced only subtle effects on performance during the continuous task of our Test trials, despite the addition of an explicit plan. Different kinematic measures contributed differently to classification between Novel and Repeat epochs for individual participants; if this represented a change in which kinematic measures the motor system minimized, we can describe that change in movement characteristics as a change in motor strategy. As an alternative explanation, participants might only have been able to minimize a few variables at a time, and that the variables involved shifted independently over time.

In addition to the motor strategy change between epochs of Test trials, we also observed between-participant differences. Examining the differences between Repeat and Novel epochs, the relative contributions of variables during each epoch were inconsistent across participants. Therefore, different participants used different motor strategies. These results are consistent with psychophysical studies of shorter three-dimensional arm movement sequences (Klein Breteler et al. 2003), in which different participants used distinct hand path patterns under identical experimental conditions.

The Test trials produced a situation with little overt performance difference between the Novel and Repeat epochs of a single trial, despite explicit knowledge of the upcoming action. Distributions of performance measures did shift, but remained largely overlapped. However, participants changed their motor strategy between Novel and Repeat epochs, with different variables providing the most information for classification under different control modes. In terms of movement generation, the addition of an internal plan differentiates untrained and trained movements; here, this added information produced statistically significant but ultimately minimal changes in distributions of error measures for reach-like movements. This contrasts sharply with learning in the SRTT (Nissen

and Bullemer 1987), wherein learning strongly affects temporal error; however, it confirms a recent preliminary report that human tracking movements involve consistently low delay times (100–130 ms) even when guided only by vision (Wessberg et al. 2006). Our results also contrast with those from another sequence task (Hikosaka et al. 1995), which demonstrated marked execution differences between learned and unlearned sequences in monkeys. However, these previous tasks all used button-pressing as a means to investigate motor sequence learning, while here we use naturalistic arm movement sequences. Any extrapolation of specific error measures or learning patterns between such markedly different motor behaviors would be difficult at best.

Because our task used externally paced movements, our kinematic observations may or may not translate to internally paced reaching movements; further studies examining this question could prove fruitful. While internally paced reaching movements may dominate normal behavior, externally paced continuous movements do have ecological relevance, in activities such as driving a car. Regardless, our central interest was to recognize when a potential neural mechanism engages for the use of an internal plan. Therefore, our results provide potentially valuable insights regardless of the specific kinematic differences between the two control modes or how neural substrates for these actions may engage.

Together, our results suggest that most but not all of the motor error enters the system "downstream" from where on-line input and internal plan information combine to guide externally paced reaching movements. This downstream source of motor noise also lies downstream from where the system selects which movement aspects (kinematic variables) to optimize in pursuit of the task goals. In other words, most motor noise is strategy-independent. We cannot determine whether visual and internal plan information combine before or during strategy generation, but the current data are inconsistent with strategy generation prior to integration of information sources. This suggests that the main sources of noise may be concomitant with very late stages in reach movement control, such as the primary motor cortex.

Classification

We used multiple classification methods to demonstrate differences between epochs. We differentiated trained and untrained manual pursuit movements most effectively using a HMM classification algorithm trained on multiple kinematic variables. The heavily overlapped distributions of kinematic error measures between control modes, combined with the failure of FLD and Bayesian algorithms to classify Catch trials, suggest that performance differed too minimally between conditions to distinguish conditions with instantaneous measures alone. However, the HMM algorithm provided at least two advantages. First, it entailed an additional structural restriction (here, that each behavioral state could only be entered once), which made it less likely to change its prediction, and therefore less likely to erroneously change its state estimate due to the high behavioral noise. Second and more importantly, it used performance history as well as current performance, allowing it to classify our data accurately even though kinematic measures overlapped so much that either trained or untrained performance could have produced most individual data samples.

These classification schemes required multiple kinematic variables to function consistently. Single kinematic variables frequently showed strong biases toward classifying data as being in one condition over the other, as demonstrated in Fig. 6. When we used five kinematic variables, these biases disappeared, which Fig. 5 reveals by the location of all points near the top-left to bottom-right diagonal (which corresponds to equal errors for each epoch). We did not analyze intermediate (2–4) numbers of variables, as the variable selection would have to differ for every participant, preventing useful generalizations.

It is important to note that we had to train our algorithms with the true task state rather than the true control mode. During the transition into the Repeat epoch, the control mode underlying the participant's behavior must have differed from the task mode not only by errors in the algorithm's plan detection, but also by the participant's reaction time and detection errors. This inevitably created a "transitional period" in which the task epoch has changed but participant performance remains unchanged. Ideally an algorithm would attempt to identify the true change in participant control mode, but the problem at hand is generating the data to train such algorithms. Because participant error (i.e., time between the change in task mode and change in participant's control mode) was inaccessible, our methods included all participant error as mistaken classifications, so the data with which we trained our algorithms can be considered only an approximation of ground truth. However, shuffling the order of the movement segments within the Repeat epoch resulted in no significant changes in reaction time, suggesting that transitional periods were trivially short compared to normal performance variability. Regardless, this limitation makes the success of our algorithms all the more remarkable, especially their ability to identify transition changes in less than 300 ms.

The low latencies for HMM identification of the epoch change require some explanation. They suggest that some aspect of the participants' sensorimotor system identified the epoch change almost immediately, even if participant

awareness normally required slightly more than one full segment to make the identification (compare 965 ms segment time with 1,099 ± 611 ms participant response time). Many sequence learning studies have demonstrated this dissociation between psychophysics and declarative knowledge (e.g., Nissen and Bullemer 1987; Willingham et al. 2002; see Ashe et al. 2006; Robertson 2007 for reviews), in which participants covertly express procedural learning before or without conscious identification of the sequence. Therefore, much of the ~855 ms delay between HMM identification and key press time probably represented a period wherein the participant could use his or her procedural knowledge of the movement sequence, without yet explicitly identifying the state change. Because of our task's novelty, we know of no previous studies addressing how long such a delay might be in a behavioral context such as the one we present. Some portion of the difference may also arise from different reaction times for pursuit tracking and key press tasks, but we did not investigate such fractionation. Participants' ability to identify the Repeat sequence almost immediately upon its start may at first seem unlikely, given the identical characteristics of Repeat-epoch segments and Novel-epoch segments. However, the Repeat sequence's position did not vary between trials, potentially allowing participants to memorize the initial segment's combination of position and direction, visually or using arm posture. Such memorization would allow participants to identify the sequence from early characteristics of its first segment, or even possibly to predict it slightly in advance as the target cursor approached that fixed sequence-initiation point.

Unlike traditional motor learning tasks such as the SRTT (Nissen and Bullemer 1987), reaching tasks allow the calculation of multiple performance measures. This allows the identification of an internal plan even when it produces few or inconsistent overt differences, as multiple subtle differences can contribute to algorithms such as the HMM method. More generally, reaching tasks allow a much more detailed analysis of performance, with multiple independent spatial and temporal domains. In the current study, this detailed analysis showed that kinematic performance between trained and untrained movements differed insufficiently to uniquely specify movement type on its own, but HMM could differentiate those movement types nonetheless. By allowing moment-to-moment classification of the information types used for movement control, these tasks and methods may allow correlation of neural recordings across motor planning areas with specific modes of motor control to determine how they generate this final performance.

Given our focus on subtle differences between movement patterns, we must recognize the limitations of our experimental apparatus. Participants had to perform a sensorimotor transformation between the movement plane defined by the tablet and the visual plane defined by the monitor. Furthermore, LCD panels have limited visual refresh rates, potentially limiting participants' ability to use proprioceptive feedback appropriately. However, our cohort of undergraduate and graduate students presumably had extensive experience with similar apparatuses, minimizing the practical differences between tasks on LCD panels and truly ecological conditions.

The possibility remains that performance changes may have arisen not from sequence learning, but from training or overtraining of specific movement segments regardless of their position in a sequence. However, recall that we used identical processes and parameters to create Repeat and Novel sequences; as such, broadly speaking, the only way to distinguish between Novel movements and Repeat movements would be by learning. Therefore, if participants improved their performance of specific segment movements, this improvement should have extended to Novel movements with similar characteristics. Our data do not allow us to directly test this, since all segment matches would be close or partial, obscuring the causes of performance changes. However, we have shown that our classification methods did not make significantly more incorrect classifications when Novel segments corresponded to the initial Repeat segment, showing that Repeat-like performance did not arise solely from individual segment movement characteristics. If they did, similar movements would have produced strongly similar performance, leading to more incorrect early identifications of Repeat movements.

While all of the movements in our task had similar characteristics, there remains at least one way in which arm dynamics could have differed between performance of Novel and Repeat movements. Some of the observed subtle kinematic differences between trained and untrained movements could have arisen from changes in limb dynamics resulting from knowledge of upcoming segments. In order to detect and measure state-based changes in dynamics, future studies might apply perturbations to the hand during reaching movements, which would reveal any changes in hand impedance.

Summary

The current work suggests that a substantial portion of motor noise enters the reach movement control system downstream from where visual and internal plan information combine. This downstream noise may explain why we did not find the expected large decreases in kinematic errors for learned movements over novel movements in this task. Additionally, we also present a continuous reaching

task with multiple error measures that may broaden the variety and complexity of tasks available for motor learning studies. Our task also demonstrates HMM as a powerful method for identifying hidden states that affect noisy variables, such the control modes used in a continuous reaching task. Finally, these features may be useful to differentiate neural populations engaged in motor performance from those that hold a plan for learned actions. Because changes in internal plan usage produce only subtle kinematic differences, brain correlates of those kinematic differences will not conceal any brain correlates of internal plan usage. While our study does not fully and conclusively elucidate the kinematic differences between learned and novel reach-like movements, it does provide the tools to identify when a computational or potential neural mechanism engages to utilize an internal plan during continuous manual movement.

Acknowledgments John P. Donoghue: NIH-NINDS NS-25074 (Javits). Jerome N. Sanes: NIH-NINDS R01NS44834.

References

Ariff G, Donchin O, Nanayakkara T, Shadmehr R (2002) A real-time state predictor in motor control: study of saccadic eye movements during unseen reaching. J Neurosci 22:7721–7729

Ashe J, Lungu OV, Basford AT, Lu X (2006) Cortical control of motor sequences. Curr Opin Neurobiol 16:213–221

Bapi RS, Doya K, Harner AM (2000) Evidence for effector independent and dependent representations and their differential time course of acquisition during motor sequence learning. Exp Brain Res 132:149–162

Bapi RS, Kiyapuram KP, Graydon FX, Doya K (2006) fMRI investigation of cortical and subcortical networks in the learning of abstract and effector-specific representations of motor sequences. NeuroImage 32:714–727

Bhat R, Sanes JN (1998) Cognitive channels computing action distance and direction. J Neurosci 18:7556–7580

Diedrichsen J, Hashambhoy Y, Rane T, Shadmehr R (2005) Neural correlates of reach errors. J Neurosci 25:9919–9931

Duhamel J-R, Colby CL, Goldberg ME (1992) The updating of the representation of visual space in parietal cortex by intended eye movements. Science 255:90–92

Geman S, Kochanek K (2001) Dynamic programming and the graphical representation of error-correcting codes. IEEE Trans Inf Theory 47:549–568

Grafton ST, Hazeltine E, Ivry RB (1998) Abstract and effector-specific representations of motor sequences identified with PET. J Neurosci 18:9420–9428

Hikosaka O, MK R, Miyachi S, Miyashita K (1995) Learning of sequential movements in the monkey: process of learning and retention of memory. J Neurophys 74:1652–1661

Karni A, Meyer G, Jezzard P, Adams M, Turner R, Ungerleider LG (1995) Functional MRI evidence for adult motor cortex plasticity during motor skill learning. Nature 377:155–158

Klein Breteler MD, Hondzinski JM, Flanders M (2003) Drawing sequences of segments in 3D: kinetic influences on arm configuration. J Neurophysiol 89:3253–3263

Lungu OV, Wächter T, Liu T, Willingham DT, Ashe J (2004) Probability detection mechanisms and motor learning. Exp Brain Res 159:135–150

Miall RC, Jackson JK (2006) Adaptation to visual feedback delays in manual tracking: evidence against the Smith Predictor model of human visually guided action. Exp Brain Res 172:77–84

Moran DW, Schwartz AB (1999) Motor cortical representation of speed and direction during reaching. J Neurophysiol 82:2676–2692

Morasso P (1981) Spatial control of arm movements. Exp Brain Res 42:223–227

Nissen MJ, Bullemer P (1987) Attentional requirements of learning: evidence from performance measures. Cogn Psych 19:1–32

Paninski L, Fellows MR, Hatsopoulos NG (2004) Spatiotemporal tuning of motor cortical neurons for hand position and velocity. J Neurophysiol 91:515–532

Poldrack RA, Sabb FW, Foerde K, Tom SM, Asarnow RF, Bookheimer SY, Knowlton BJ (2005) The neural correlates of motor skill automaticity. J Neurosci 25:5356–5364

Press SJ (1989) Bayesian statistics: principles models and applications. Wiley, New York

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286

Robertson EM (2007) The serial reaction time task: implicit motor skill learning? J Neurosci 27:10073–10075

Sakai K, Hikosaka O, Miyauchi S, Takino R, Sasaki Y, Pütz B (1998) Transition of brain activation from frontal to parietal areas in visuomotor sequence learning. J Neurosci 18:1827–1840

Shadmehr R, Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. J Neurosci 14:3208–3224

Shadmehr R, Wise SP (2005) The computational biology of reaching and pointing: a foundation for motor learning. MIT Press, Cambridge

Wang T, Dordevic GS, Shadmehr R (2001) Learning the dynamics of reaching movements results in the modification of arm impedance and long-latency perturbation responses. Biol Cybern 85:437–448

Wang W, Chan SS, Heldman DA, Moran DW (2007) Motor cortical representation of position and velocity during reaching. J Neurophysiol 97:4258–4270

Wessberg J, Perfiliev S, Fridlund M (2006) Continuous arm guidance by a visual target during high-speed tracking. Program No. 440.11. 2006 Neuroscience Meeting Planner. Society for Neuroscience, Atlanta, GA (online)

Willingham DB, Salidis J, Gabrieli JDE (2002) Direct comparison of neural systems mediating conscious and unconscious skill learning. J Neurophysiol 88:1451–1460