

Dynamic Estimation of Task-Relevant Variance in Movement Under Risk

Michael S. Landy Julia Trommershäuser Nathaniel D. Daw

Department of Psychology and Center for Neural Science
New York University, New York, NY

Abstract: 174 words (250 max)

Introduction: 500 words (500 max)

Discussion: 874 words (1500 max)

Running head: Variance estimation under risk

Corresponding author:

Michael Landy
Department of Psychology
New York University
6 Washington Place, Room 961
New York, NY 10003
USA

Tel: (212) 998-7857
Fax: (212) 995-4349
Email: landy@nyu.edu

Abstract

Humans take into account their own movement variability as well as potential consequences of different movement outcomes in planning movement trajectories. When variability increases, planned movements are altered so as to optimize expected consequences of the movement. Past research has focused on the steady-state responses to changing conditions of movement under risk. Here, we study the dynamics of such strategy adjustment in a visuomotor decision task in which subjects reach toward a display with regions that lead to rewards and penalties, under conditions of changing uncertainty. In typical reinforcement-learning tasks, subjects should base subsequent strategy by computing an estimate of the mean outcome (e.g., reward) in recent trials. In contrast, in our task strategy should be based on a dynamic estimate of recent outcome uncertainty (i.e., squared error). We find that subjects respond to increased movement uncertainty by aiming movements more conservatively with respect to penalty regions, and that the estimate of uncertainty they use is well characterized by a weighted average of recent squared errors, with higher weights given to more recent trials.

Keywords: hand movements, movement planning, uncertainty estimation, visuomotor control

Introduction

Humans take uncertainty into account when planning and executing movements, including movements around obstacles (Sabes and Jordan, 1997; Sabes et al., 1998) or toward targets whose location is uncertain (Körding and Wolpert, 2004; Tassinari et al., 2006). Humans also take visuomotor uncertainty into account when planning a fast reaching movement under risk (Trommershäuser et al., 2003b, a; Trommershäuser et al., 2006; Maloney et al., 2007). In these experiments, humans reach under a time constraint at a visual display consisting of a target region and neighboring penalty region. Hits on the target that meet the time constraint resulted in financial rewards and hits on the penalty can result in a net loss for that trial. Characterizing movement selection as a gain-maximization problem, human performance is nearly optimal in these tasks. Humans also change movement strategy in response to changing uncertainty, for example when movement uncertainty is changed artificially in a virtual environment (Trommershäuser et al., 2005), using a visuo-motor reflex (Hudson et al., 2010), or is naturally larger in one direction than another (Gepshtein et al., 2007). Humans also perform well in a task in which they choose when to move so as to achieve a particular tradeoff of motor and visual uncertainty (Battaglia and Schrater, 2007).

However, the foregoing work addresses steady-state behavior and leaves open many questions about what sort of dynamic adjustment or learning drives these changes. This is contrasting (and complementary) to a different body of work on gain optimization in more abstract decision tasks, such as “bandit” tasks

(Bayer and Glimcher, 2005; Sugrue et al., 2005; Daw et al., 2006; Behrens et al., 2007). There, participants' choices and associated neural signals are well characterized by a process that dynamically tracks an estimate of the mean gain expected for an option, as embodied by reinforcement learning models using delta rules or reward gradients (Sutton and Barto, 1998). The role of uncertainty in reinforcement learning tasks has been a question of sustained interest also, but has been relatively difficult to address experimentally since in these tasks uncertainty about the mean payoff tends to contribute only indirectly to behavior, such as by controlling the stepsizes for updates (Courville et al., 2006; Behrens et al., 2007; Li et al., 2011) or the degree of exploration (Daw et al., 2006; Frank et al., 2009).

In this study, we use methods from the reinforcement learning literature to study dynamic strategic adjustment in a visuomotor decision task in which subjects must cope with movement consequences whose uncertainty changes over time. Since in our task, the gain-optimizing strategy is determined by the uncertainty (i.e., variance) in the movement endpoint, rather than its mean, this offers an opportunity to examine how humans learn second-order statistics from feedback when these statistics have relatively direct behavioral consequences. Indeed, the straightforward application of a standard mean-estimating reinforcement rule to our task predicts a qualitatively and quantitatively different pattern of results, and thus offers a crucial comparison point for ensuring we are isolating behavioral adjustments related to changing uncertainty.

Methods

Stimuli and procedure

Subjects performed a rapid reaching movement and attempted to hit a target region displayed on a touchscreen while, on a subset of trials, avoiding a partially overlapping penalty region (Fig. 1). Subjects began each trial with their right index finger pressing the CTRL key of a keyboard firmly attached to a table. A fixation cross was displayed for 1 s, followed by the target display. Subjects could begin the reach at any time after stimulus display, but were constrained to complete the movement within 400 ms.

Figure 1 about here

The target was a green outlined rectangle, 182×64 mm, located 37 mm left or right of the center of the display, viewed from a distance of 40 cm. On target-only trials, only the green rectangle was displayed. On penalty trials, an overlapping solid red penalty rectangle of the same size was displayed, displaced left or right of the target rectangle by half its width (Fig. 1A). The display background was black.

Task-relevant variability was manipulated by adding a horizontal perturbation to the movement endpoint. When the finger reached the screen within the time constraint, two small squares were displayed: a white square at the actual endpoint of the movement (with horizontal position x_{hit}), and a blue square (with

horizontal position x_{fb}) shifted horizontally from that position by a random amount:

$$x_{fb} = x_{hit} + \Delta \text{ (Fig. 1B).}$$

The movement perturbation Δ was normally distributed with zero mean and a standard deviation (SD), σ_{pert} , that could change from trial to trial. σ_{pert} followed a “sample-and-hold” trajectory. The initial value of σ_{pert} was chosen randomly and uniformly from a range of 3.7 to 18.4 mm. σ_{pert} remained constant for an epoch that lasted for 75 to 150 trials; the epoch length was chosen randomly and uniformly over that range. At the end of the epoch a new value of σ_{pert} and epoch length were chosen similarly, and so on until the end of the block of trials.

Subjects participated in one training and two experimental sessions. The training session consisted of 300 target-only trials. Experimental sessions consisted of 420 trials, including 140 target-only trials and 280 penalty trials, with equal numbers of trials with the target on the left or right, and for penalty trials, equal numbers of trials with the penalty region to the left or right of the target. Both training and experimental sessions began with an additional 12 warmup target-only trials; the data from the warmup trials were not analyzed.

Reward feedback was based on the position x_{fb} of the shifted, blue square. If the square fell within the target region, the subject earned one point (4¢). If it fell within the penalty region, they lost 5 points. In the overlap region, both the reward and penalty were awarded. Movements that failed to meet the time constraint resulted in a 10-point penalty. If subjects began the movement before the stimulus was displayed, the trial was re-run. Subjects were given feedback at

the end of every trial with the results of that trial, and were told their cumulative score at the end of each session.

Apparatus

The experiment was controlled by a PC and stimuli were presented on an ELO Touchsystems (Menlo Park, CA) ET1726C 17-in CRT display with touchscreen, with claimed positional measurement accuracy SD of 2 mm. Stimulus timing and display were controlled using the Psychophysics Toolbox package (Brainard, 1997; Pelli, 1997)

Subjects

Seven subjects ran in the experiment including one author (MSL). All had normal or corrected-to-normal vision. All but one (JMF) were right-handed and all used the right hand to perform the reaches. Subjects signed a consent form approved by the NYU University Committee on Activities Involving Human Subjects. They were paid a base fee plus a bonus based on performance in the task. Bonuses ranged from \$2.80 to \$14.44 for a single practice or experimental session.

Ideal-performance model

We treat ideal performance as the choice of an aimpoint that results in maximum expected gain. Such a strategy must take into account the stimulus display (the locations of payoff and penalty regions), the reward and penalty values, as well as the total task-relevant uncertainty. For our task, both the target

and penalty regions are large vertical rectangles, so tall as to be impossible to miss in the vertical direction. Thus, this is effectively a one-dimensional task, and we need only model the choice of horizontal aimpoint x_{aim} . On any given trial, movement uncertainty has two components: the subject's own motor uncertainty σ_{motor} and the experimenter-imposed outcome uncertainty σ_{pert} , resulting in overall uncertainty $\sigma_{\text{overall}} = \sqrt{\sigma_{\text{pert}}^2 + \sigma_{\text{motor}}^2}$. In reaching tasks such as these, we find endpoint distributions to be well approximated by a Gaussian distribution (Trommershäuser et al., 2005) and so, for the purposes of modeling ideal behavior in this task, we assume that when a participant aims at location x , endpoints will be distributed as a Gaussian with variance $\sigma_{\text{overall}}^2$.

If the participant had precise knowledge of σ_{overall} , then they could select an ideal strategy for the task, i.e., choose an aimpoint that maximized expected gain, providing the best tradeoff between maximizing hits on the target while minimizing occasional hits on the penalty region. Figure 2 illustrates the computation of the optimal aimpoint x_{opt} for this “omniscient” observer who somehow has knowledge of the value of σ_{overall} for the current trial. Figure 2A shows the distribution of movement feedback locations x_{fb} resulting from a particular value of x_{aim} (indicated by the vertical black line through all three panels). The experimenter-imposed gain function is shown in Figure 2B, with the corresponding probabilities of hitting the target or penalty indicated by the black and gray cross-hatching in Figure 2A, respectively. The expected gain for this

aimpoint is indicated by the black square in Figure 2C, and is computed as the sum of the target and penalty values weighted by the respective probabilities:

$$EG(x_{\text{aim}}) = 1 \times P(\text{hit-target} | x_{\text{aim}}, \sigma_{\text{overall}}) - 5 \times P(\text{hit-penalty} | x_{\text{aim}}, \sigma_{\text{overall}}). \quad (1)$$

Figure 2C shows expected gain as a function of aimpoint. The aimpoint corresponding to the peak of the curve (the black circle) is the optimal strategy x_{opt} .

 Figure 2 about here

Figure 3 shows the optimal strategy x_{opt} as a function of overall noise variance. As one might expect, increased noise leads an optimal omniscient participant to aim further away from the penalty region. In this figure, the vertical lines bound the region of $\sigma_{\text{overall}}^2$ values that occurred in our study (across participants and epochs).

 Figure 3 about here

Analyses of dynamic learning

We are interested in determining how strategy on a given trial is determined based on the outcomes of previous trials. In our task, an unexpectedly large perturbation should drive a (variance-estimating, gain optimizing) participant to shift his or her aim further away from the penalty area, regardless of whether that perturbation drove the reinforced location x_{fb} toward or away from the penalty. A strategy based on tracking the mean perturbation, analogous to a reinforcement

learner, will instead adjust aimpoints to compensate for the average perturbation: *reducing* the shift (away from the penalty) after trials with large perturbations that move the reinforced location x_{fb} further away from the penalty. (Because the position of the penalty is randomly to the left or right of the target, we conduct all our analyses using coordinates for x in which positive values indicate the direction away from the penalty region, rather than left or right.)

To contrast these two approaches, and to verify that aim shifts were related to variance rather than mean tracking, we will compare strategies that compute aimpoints based on the average of recently experienced *squared* perturbation (an estimator for the variance) to those that track the average of the perturbations themselves (the mean). We consider models that compute the aimpoint x_{aim} as a linear function of the statistics of recently experienced perturbations. As an approximation to the ideal model, this tractable approach is justified in that within the range of values of $\sigma_{overall}^2$ that confronted participants, the optimal aimpoint was approximately a linear function of $\sigma_{overall}^2$ (Fig. 3), and hence of σ_{pert}^2 as well.

Regression analyses

We will begin by examining a particularly simple model in which the current aimpoint is based solely on the most recently experienced perturbation. For such a prediction to be made, the current trial must be a penalty trial (participants should aim at the center of the target on target-only trials). Thus, we will regress the current reach endpoint on penalty trials against either the previously

experienced perturbation or squared perturbation. Participants only experience the perturbation on trials in which the finger arrived at the display screen prior to the timeout; for trials in which the reach was too slow, the perturbed location was not displayed. In subsequent analyses, we will examine models that base the strategy on more than one previously experienced perturbation (on more than one previous trial). To compare these models on an equal basis, all models are fit only to penalty trials in which there were at least 15 previously experienced perturbations in the same session (excluding warmup trials).

We will also determine whether current strategy is based on a longer history than merely the most recently experienced perturbation. To do so, we will conduct “lagged regressions” where, instead of using only the most recently experienced perturbation or squared perturbation as a regressor, we will include the 8 most recently experienced perturbations or squared perturbations as regressors.

Model fits

We will also consider parametric models that, like the lagged regression, base the current trial’s strategy on a weighted sum of recently experienced perturbation. These models, like the results of the lagged regression shown below, give the greatest weight to the most recent trials, by using weights that decay exponentially with the lag. We will compare several different models of this form to ask about the relative importance of variance- and mean-estimation in subjects’ choices. We begin by describing the most complex model we fit to the data; the other models are nested simplifications of this model.

The models assume that the aimpoint in each penalty trial is a linear combination of previously experienced perturbations and/or squared perturbations. Participants experience a perturbation (i.e., see a display indicating the actual and perturbed endpoints as in Fig. 1B) on any trial (penalty or target-only) in which the reach arrived at the display screen within the experimenter-imposed time constraint. We index these “useable” trials (not a warmup trial, no timeout) by $i = 1, 2, \dots, N$, where N is the total number of useable trials in that experimental session. The regression is computed to model data only for those useable trials j meeting two additional constraints: j is a penalty trial and $j > 15$ (so there are enough previous trials on which to base the prediction). For such a trial j , the model for the reach endpoint is

$$x_{\text{hit}j} = B + w_s \sum_{i=1}^{j-1} e^{-(j-i-1)/\tau_s} \Delta_i^2 + w_p \sum_{i=1}^{j-1} e^{-(j-i-1)/\tau_p} \Delta_i + \varepsilon_j. \quad (2)$$

Here, Δ_i is the value of the perturbation in useable trial i , where positive values indicate a perturbation in the direction away from the penalty region (i.e., rightward when the penalty was to the left of the target, and leftward otherwise). w_s and τ_s are the weight and exponential time constant for squared perturbations, and w_p and τ_p are the weight and time constant for perturbations.¹

¹ Early trials in the session have fewer previous perturbations to contribute to the model’s prediction of the current trial’s value of x_{hit} . We also fit models in

which the weight on Δ_i^2 were normalized by $\sum_{i=1}^{j-1} e^{-(j-i-1)/\tau_s}$ (and similarly for the

B is a bias term; think of it as the mean shift in the absence of any information about the current perturbation. Finally, reach aimpoints are perturbed by motor noise ε_j that has zero mean, is normally distributed, independent over trials, and has standard deviation σ_{motor} . Thus, there are five parameters modeling the aimpoint (B , w_s , τ_s , w_p and τ_p) and one noise parameter σ_{motor} . The estimated noise parameter is simply the standard deviation of the residuals. The models were fit to the data of both experimental sessions by maximum likelihood, separately for each subject.

We will compare seven different models:

1. “Exponential mean and variance” described by Eq. 2,
2. “Exponential variance only” that omits the terms based on previously experienced perturbations (i.e., forcing $w_p = 0$),
3. “Exponential mean only” that omits the terms based on previously experienced squared perturbations (i.e., forcing $w_s = 0$),
4. “1-back variance and mean” that only considers one previously experienced trial (i.e., forcing τ_s and τ_p to be vanishingly small),
5. “1-back variance only” that considers one previously experienced squared perturbation (i.e., small τ_s and $w_p = 0$),

weights on Δ_i). These models behaved similarly to the unnormalized versions summarized here.

6. “1-back mean only” that considers one previously experienced perturbation (i.e., small τ_p and $w_s = 0$), and
7. “Shift bias alone” (i.e., $w_s = w_p = 0$).

Each of these models was fit to the data for each subject separately by maximum likelihood. Models 5 and 6 are formally identical to the linear described in the previous subsection, and Model 4 is an analogous bivariate linear regression. Many pairs of models are nested in the sense that one model is a constrained version of the other, and thus may be compared using the nested hypothesis test (Mood et al., 1973, pp. 440-442).

Results

Raw endpoint data

Results for two individual subjects are shown in Figure 4. For each subject, panel A shows the trajectory of imposed outcome uncertainty (σ_{pert}) over the two experimental sessions. We estimated each subject’s motor uncertainty, σ_{motor} , as the standard deviation of the horizontal location of movement endpoints relative to the center of the target, pooled across all trials in the training session and all no-penalty trials in the two experimental sessions (excluding any trials in which the participant failed to arrive at the target in time). Overall uncertainty was estimated as the combination of these two independent sources of uncertainty, as noted before (see *Ideal-performance model*).

Figure 4 about here

Comparison to the optimal “omniscient” strategy

Figure 4B shows the optimal aimpoints for each penalty trial in the experimental sessions (the solid and dashed lines) as well as the actual endpoints for all penalty trials in which the reach arrived at the screen prior to the timeout. A y-axis value of zero corresponds to the center of the target, and positive values correspond to landing points shifted away from the penalty, allowing us to sensibly plot data together for trials in which the penalty region was to the left or right of the target. The plotted optimal aimpoints are based on each trial’s value of σ_{overall} (see Fig. 3), which, in turn, depends on the value of σ_{pert} , which was not available to the participants. Hence, we refer to this as the optimal “omniscient” strategy. Nonetheless, there is some indication in Figure 4B, especially for PSH, that the participant’s strategy was affected by changing uncertainty; endpoints indicate a greater degree of caution (landing further from the penalty region) during periods of greater outcome uncertainty.

We determined the degree to which participants compensated for changing uncertainty by comparing reach endpoints to those predicted by the optimal omniscient strategy. Figure 5A plots reach endpoints for one subject as a function of optimal omniscient endpoints. A linear regression (dashed line) indicates that there was substantial compensation for changing uncertainty by this subject. The amount of compensation in response to the applied perturbation

varied across subjects (Fig. 5B), and was significant for six of the seven participants. Most subjects compensated less than called for by the optimal omniscient strategy, but one subject (SO) overcompensated.

Figure 5 about here

Response to recently experienced perturbations

In the preceding analysis, participants' strategies of aimpoint selection were compared to the omniscient strategy, which assumes the participant has access to the current trial's value of σ_{pert} . Of course, although aimpoints were sensitive to this parameter, its value was not told to the participants, and furthermore was changed over trials. Thus this sensitivity must have arisen by participants adjusting their strategy in light of experience, e.g. by dynamically estimating outcome uncertainties to formulate their aimpoints on each trial.

Our remaining analyses seek to characterize these dynamic adjustments by studying the dependence of aimpoints on recent experience. Rather than specifying a full Bayesian ideal updater (Behrens et al., 2007), we seek to determine what statistics of recent observations influence the aim strategy, and in so doing, to probe a key difference in the present task from many reinforcement learning tasks that have been studied with similar analyses (Corrado et al., 2005; Lau and Glimcher, 2005). Gain optimization in reward learning tasks is typically driven by estimating mean payoffs (e.g., with delta-rule or gradient-climbing algorithms; Sutton and Barto, 1998; Dayan and Abbott, 2005), so as to choose the option with highest rewards. However, in the present

task, the key statistic determining the optimal strategy is the variance of the perturbations, not their mean.

Figure 6 illustrates the results of regressing the current reach endpoint on penalty trials against either the previously experienced squared perturbation (panels *A* and *B*) or the previously experienced perturbation (panels *C* and *D*). The influence of the previously experienced squared perturbation on the current trial's reach endpoint is significant averaged across observers ($t(6) = 2.63$, 1-tailed $p = 0.02$), and is significant individually for 6 out of 7 participants (Fig. 6A), although the degree of influence (Fig. 6A) and variance accounted for by this regression (Fig. 6B) differ substantially across participants. The regression coefficients are all positive, indicating that larger squared perturbations lead participants to aim further away from the penalty, as expected. In contrast, the previously experienced perturbation (not squared) does not have a significant influence on the current reach endpoint averaged across observers ($t(6) = -1.1$, 1-tailed $p = 0.15$). This influence is not significant for any individual observer and the signs of the regression coefficients are inconsistent across subjects (Fig. 6C).

Figure 6 about here

Next, to study how participants combined perturbations from multiple trials into an estimate, while making minimal parametric assumptions about the form of this combination, we conducted a similar regression where, instead of using only the most recently experienced squared perturbation as a regressor, we included

the 8 most recently experienced squared perturbations as regressors. The weights on each recent perturbation vary substantially across subjects (Fig. 7). The average influence of perturbations differs significantly from zero for 5 of the 8 lags ($p < .05$), and approaches significance for the other 3 lags. The overall trend is noisy, but appears to weight more recent perturbations more heavily than those further in the past.

This decay appears, at least on the average (which is an estimator for the population-level effect; Holmes and Friston, 1998) roughly exponential in shape. Such an exponential form to the lagged dependence is predicted if subjects track their variance estimate by an error-driven running average (here, of squared perturbations; e.g., Preuschoff et al., 2008).

Figure 7 about here

Parametric models

The results of the lagged regression indicate that participants base their choice of reach aimpoints on several recently experienced perturbations, with greater weight placed on more recent trials. However, characterizing the lagged dependence given so few structural assumptions about its form requires a large number of free parameters and introduces concerns of overfitting. Therefore, in order to directly capture the effects of squared and non-squared perturbations together in a single model, and to characterize this dependence over arbitrary timelags, we next adopted the parameterization suggested by the previous

analysis. That is, we considered models that weight recently experienced perturbations using weights that decay exponentially with lag.

First, we tested whether this approach was justified by comparing the fit to data of the unconstrained model from Figure 7, to the corresponding model assuming exponentially decaying weights (Model 2). Although the latter has many fewer free parameters, it fit the data nearly as well. In particular, penalizing the fit for the free parameters using either the AIC or BIC criterion (because the two models are not nested; Schwarz, 1978), the exponential model fit better for every participant, by an average AIC difference of 108 and BIC difference of 134.

Figure 8 presents the results of the model comparisons. Each rectangle corresponds to one of the 7 models, and gives the number of model parameters (excluding the noise parameter σ_{motor} required by every model). Arrows indicate nesting between models, with more complex models placed higher in the figure. Next to each arrow is an indication of the number of subjects of the 7 for whom the hypothesis test rejected the null hypothesis ($p < .05$, with no corrections for multiple tests) indicating that the extra parameters in the more complex model were justified.

Figure 8 about here

We can summarize the model comparison tests as follows. First, in every case in which the more complex model of the pair adds squared perturbations as a predictor (3 model comparisons), the model comparison supports the addition of parameters for 6 out of 7 subjects. Second, for any 1-back model that includes

predictions based on squared perturbations, model comparisons justify the addition of an exponential weighting of past trials for 6 out of 7 subjects (there are 2 such model comparisons). Finally, the evidence is somewhat weaker for the usefulness of including nonsquared perturbations for prediction. Model comparisons that add in these regressors are not significant for any subjects for the 1-back models (there are 2 such model comparisons). But, adding in an exponential weighting of nonsquared regressors is significant for 5 out of 7 subjects (there are 2 such model comparisons). Thus, combining all of these observations, the evidence for Model 2 (Exponential variance only) is strongly supported by the data, indicating that participants base their aimpoints on an estimate of perturbation variance based on an exponentially weighted average of previously experienced squared perturbations. But, there is some evidence for a response to an exponentially weighted average of past perturbations. We should note, however, that Model 1, which also includes perturbations for prediction, explains very little additional variance as compared to Model 2 for most subjects; in the median across subjects, the variance accounted for by Model 1 is only a factor of 1.113 larger than that accounted for by Model 2.

How does Model 2, which uses only squared perturbations, compare to Model 3, which uses only raw perturbations as regressors? Although the relative fits of these two models to the data cannot be tested using classical methods, since they are not nested, the two models do contain the same number of free parameters so their raw goodness of fit (e.g., in terms of variance accounted for) can be fairly compared. Excluding subject AI, for whom neither model explains

an appreciable fraction of the variance, for four of the subjects, Model 2 fares much better. For all 6 other subjects, the fit of Model 2 is preferred in terms of log likelihood (equivalent to an AIC or BIC calculation, since the numbers of parameters are identical). For these subjects, the ratio of variance accounted for (Model 3 relative to Model 2) ranges from .09–.18. However, for two subjects (JF and NGC), Model 3 explains nearly as much variance as Model 3 (ratios of .68 and .72, respectively). Thus, for these two subjects especially, a series of perturbations away from the penalty resulted in responses closer to the penalty in subsequent trials (and vice versa) in addition to the predicted response to large (squared perturbations) driving responses away from the penalty.

The models involving responses to the mean perturbation use a coordinate system for both the perturbations and responses in which positive values are for perturbations or movement endpoints shifted away from the penalty. In the motor-adaptation literature, one typical scenario is to study adaptation to a particular perturbation of all movements (say, 3 cm rightward), regardless of movement direction or endpoint. We have also performed model comparisons on our data of this sort, looking for *leftward* shifts of endpoint in response to recent *rightward* perturbations in the models involving response to mean perturbation. In these model comparisons, the models involving response to the mean have far less support, showing significant results for only 2 out of 7 subjects (compared to 5 subjects as noted above) and for this coding of perturbation, the ratio of variance accounted for by Model 1 compared to Model 2 is only 1.016. Thus, we

have no evidence for response to the mean perturbation coded relative to world (rather than target-penalty) coordinates.

Figure 9 compares the lagged regression results from Figure 7 with the corresponding exponential fits from Model 2. Figure 9A shows the lagged regression weights on the 8 previously experienced squared perturbations (solid lines) and the corresponding exponential fits of Model 2 (dashed lines) separately for each subject. The correspondence is reasonable for the subjects with the largest weights (note that the y -axis scales vary across subjects). For the others, the correspondence seems weaker. The reason for this can be seen more clearly in a scatterplot of the weights fit by both techniques (Fig. 9B). For several subjects, the weights from the lagged regression vary considerably compared to those huddled just above zero from the exponential fit of Model 2. The reason is that the best-fit time constants (τ_s) were often large (ranging from 2.2 to 217 trials), whereas the lagged regression was constrained to explain the data using only the 8 most recently experienced perturbations.

Figure 9 about here

In summary, we have demonstrated that humans change the movement plan for a reach in response to experimenter-imposed changes in task-related variance in movement under risk. In this task with explicit perturbation of reinforced movement outcome, participants' strategies appear to be suboptimal, and vary substantially over participants. Nevertheless, almost all participants responded to changes in the variance of the perturbation, shifting their aimpoint

further away from the penalty region when perturbation variance was increased. However, participants varied substantially in both the extent of shift in response to changes in movement outcome variability as well as the time constant of the exponential average of past squared perturbations that controlled that shift. Some subjects also varied strategy in response to the mean perturbation, but this effect was, for most subjects, far smaller.

Discussion

This experiment aimed at exploring how humans dynamically estimate and respond to variability in the outcome of movement under risk. Participants performed rapid reaches at a target while attempting to avoid a neighboring penalty region. The experimenter-imposed variability was much larger than motor variability, was repeatedly changed, and was visible to participants because the actual and perturbed reach endpoints were displayed simultaneously. Most participants responded to increased movement outcome uncertainty with greater caution by aiming further away from the penalty region. We characterized the trial-by-trial dynamics of this adjustment, and distinguished it from other potentially confounding adjustments related to trial-by-trial tracking of and compensation for the mean perturbation.

Descriptively, our data appear to be well characterized by a model in which subjects track variance by an exponentially-weighted average over past squared perturbations. This is a mechanistically plausible model. Unlike ideal-observer Bayesian estimators, which for nontrivial change processes can typically be solved only numerically (e.g., Behrens et al., 2007), such a weighting is easy to

compute online. In particular, an exponential weighting results from maintaining only a running average over squared perturbations, and at each trial adjusting this by a delta rule to reduce the difference between the obtained squared perturbation and the current estimate: in effect, the prediction error for the (squared) prediction error. In reinforcement learning models, the same rule for mean-tracking (i.e., the prediction error rule for learning mean, non-squared payoffs) is the basis for the familiar Rescorla-Wagner (1972) and temporal-difference learning (e.g., Montague et al., 1996; Sutton and Barto, 1998) models.

Accordingly, in studies of discrete decision tasks for reward, a result analogous to Figure 7 for the lagged (not squared) rewards has repeatedly been reported, and has there been interpreted as evidence for such a learning rule (Bayer and Glimcher, 2005; Corrado et al., 2005; Lau and Glimcher, 2005). The suggestion that variances might also be estimated by a similar running-average rule of the squared errors has roots as far back as the model of Pearce and Hall (1980), at least when re-interpreted in statistical terms (Sutton, 1992; Dayan and Long, 1998; Dayan et al., 2000; Courville et al., 2006), and neural correlates for variance estimates, errors in them, or other related quantities have also been reported (Behrens et al., 2007; Preuschoff et al., 2008; Li et al., 2011). However, because in reinforcement learning tasks, the uncertainty affects behavior only indirectly, the timecourse of variance estimation has not, to our knowledge, previously been directly estimated or visualized as a function of lagged observations in the way reported here.

A related point is that whereas the delta rule for mean tracking has a well understood relationship with an ideal-observer model (in particular, it arises asymptotically in the Kalman filter), it remains to be understood the extent to which similar weighted averaging rules approximate the variance estimates that would arise for Bayesian variance estimation, and if so, for what change process. However, one hallmark of exact Bayesian change tracking across many change processes (DeWeese and Zador, 1998), but which does not hold in our approximate model, is that it is much easier to detect an increase as compared to a decrease in variability. This is because a single outlier is strong evidence for an increase in variance, whereas a sequence of near-mean observations does not guarantee that variance has decreased. Thus, since in the current experiment, movement outcome variability can suddenly increase or decrease within a session, if participants are engaging in Bayesian variance tracking, one might expect a shorter time constant for adaptation of reach strategy to increases in variance than to decreases. We have looked for and failed to find this in the present data, although the variance trajectories we used were not designed for this analysis.

In previous work, we have often demonstrated optimal behavior in reaching tasks similar to this one, indicating that humans can take into account their own movement uncertainty in planning reaches under risk (Trommershäuser et al., 2003b, a), even when that uncertainty is altered surreptitiously by the experimenter (Trommershäuser et al., 2005; Hudson et al., 2010). And, an estimate of perceptual and motor uncertainty is used in a manner reasonably

consistent with predictions of a Kalman filter in adapting to average error in motor tasks (Baddeley et al., 2003; Burge et al., 2008).

In the current experiment, although we have not computed an ideal-performance model for our experimental conditions, the inter-subject variability makes it clear that performance is not near-optimal for all participants. Our results indicate that optimal performance is mostly achieved in reaching-under-risk experiments in which the conditions are reasonably natural (Trommershäuser et al., 2003b, a; Trommershäuser et al., 2005; Hudson et al., 2010). When the task is slow and deliberative (Landy et al., 2007) or the feedback conditions involve an obviously artificial perturbation as they do here, performance becomes suboptimal and large inter-subject differences arise. We also note that optimally tracking a changing quantity, and then dynamically responding optimally with respect to these estimates trial-by-trial, is a qualitatively different and arguably more difficult criterion than making optimal decisions in the steady state. Nevertheless, what is clear from our analysis is that humans can and do use a dynamic estimate of current uncertainty, based on recently experienced errors, to plan reaches under risk.

Acknowledgments

This research was supported by National Institutes of Health Grant EY08266; JT was funded by the Deutsche Forschungsgemeinschaft (DFG/Heisenberg-Program).

References

- Baddeley RJ, Ingram HA, Miall RC (2003) System identification applied to a visuomotor task: near-optimal human performance in a noisy changing task. *J Neurosci* 23:3066-3075.
- Battaglia PW, Schrater PR (2007) Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. *J Neurosci* 27:6984-6994.
- Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47:129-141.
- Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10:1214-1221.
- Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10:433-436.
- Burge J, Ernst MO, Banks MS (2008) The statistical determinants of adaptation rate in human reaching. *J Vis* 8:20 21-19.
- Corrado GS, Sugrue LP, Seung HS, Newsome WT (2005) Linear-Nonlinear-Poisson models of primate choice dynamics. *Journal of the experimental analysis of behavior* 84:581-617.
- Courville AC, Daw ND, Touretzky DS (2006) Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences* 10:294-300.
- Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441:876-879.

- Dayan P, Long T (1998) Statistical models of conditioning. In: Advances in Neural Information Processing Systems 10 (Kearns MJ, Jordan MI, Solla SA, eds), pp 117-123. Cambridge, Mass.: MIT Press.
- Dayan P, Abbott LF (2005) Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Cambridge, Mass.: MIT Press.
- Dayan P, Kakade S, Montague PR (2000) Learning and selective attention. *Nat Neurosci* 3 Suppl:1218-1223.
- DeWeese M, Zador A (1998) Asymmetric dynamics in optimal variance adaptation. *Neural Comput* 10:1179-1202.
- Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat Neurosci* 12:1062-1068.
- Gepshtein S, Seydell A, Trommershauser J (2007) Optimality of human movement under natural variations of visual-motor uncertainty. *Journal of vision* 7:13:11-18.
- Holmes AP, Friston KJ (1998) Generalisability, random effects and population inference. *NeuroImage* 7:S754.
- Hudson TE, Tassinari H, Landy MS (2010) Compensation for changing motor uncertainty. *PLoS Comput Biol* 6:e1000982.
- Körding KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. *Nature* 427:244-247.
- Landy MS, Goutcher R, Trommershäuser J, Mamassian P (2007) Visual estimation under risk. *Journal of vision* 7:4.

- Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior* 84:555-579.
- Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND (2011) Differential roles of human striatum and amygdala in associative learning. *Nat Neurosci* 14:1250-1252.
- Maloney LT, Trommershäuser J, Landy MS (2007) Questions without words: A comparison between decision making under risk and movement planning under risk. In: *Integrated Models of Cognitive Systems* (Gray W, ed), pp 297-315. New York: Oxford University Press.
- Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936-1947.
- Mood AM, Boes DC, Graybill FA (1973) *Introduction to the theory of statistics*, 3d Edition. New York,: McGraw-Hill.
- Pearce JM, Hall G (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review* 87:532-552.
- Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis* 10:437-442.
- Preuschoff K, Quartz SR, Bossaerts P (2008) Human insula activation reflects risk prediction errors as well as risk. *J Neurosci* 28:2745-2752.

- Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In: Classical Conditioning, 2: Current Research and Theory (Black AH, Prokasy WF, eds), pp 64-69. New York: Appleton Century-Crofts.
- Sabes PN, Jordan MI (1997) Obstacle avoidance and a perturbation sensitivity model for motor planning. *J Neurosci* 17:7119-7128.
- Sabes PN, Jordan MI, Wolpert DM (1998) The role of inertial sensitivity in motor planning. *J Neurosci* 18:5948-5957.
- Schwarz GE (1978) Estimating the dimension of a model. *Annals of Statistics* 6:461-464.
- Sugrue LP, Corrado GS, Newsome WT (2005) Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature reviews Neuroscience* 6:363-375.
- Sutton RS (1992) Gain adaptation beats least squares? In: Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems, pp 161-166. Yale University, New Haven, Conn.
- Sutton RS, Barto AG (1998) Reinforcement learning an introduction. In: Adaptive computation and machine learning. Cambridge, Mass.: MIT Press.
- Tassinari H, Hudson TE, Landy MS (2006) Combining priors and noisy visual cues in a rapid pointing task. *J Neurosci* 26:10154-10163.
- Trommershäuser J, Maloney LT, Landy MS (2003a) Statistical decision theory and trade-offs in the control of motor response. *Spat Vis* 16:255-275.

Trommershäuser J, Maloney LT, Landy MS (2003b) Statistical decision theory and the selection of rapid, goal-directed movements. *J Opt Soc Am A Opt Image Sci Vis* 20:1419-1433.

Trommershäuser J, Landy MS, Maloney LT (2006) Humans rapidly estimate expected gain in movement planning. *Psychol Sci* 17:981-988.

Trommershäuser J, Gepshtein S, Maloney LT, Landy MS, Banks MS (2005) Optimal compensation for changes in task-relevant movement variability. *J Neurosci* 25:7169-7178.

Figure Captions

Figure 1. Stimuli and task. **A**, The initial stimulus display consisted of a green outlined target region and an overlapping red penalty region (here shown as outlined black and solid gray, respectively). Participants performed a speeded reach with the right hand, attempting to hit the target while avoiding the penalty region. **B**, After the movement was complete, a small white square indicated the reach endpoint. A small blue square was also displayed, randomly displaced horizontally away from the reach endpoint (here shown as open and filled squares, respectively). The participant received 1 point (4¢) for each endpoint within the target region and lost 5 points for hitting the penalty region. Slow movements (> 400 ms) led to a 10-point penalty.

Figure 2. Expected gain and optimal strategy. **A**, Illustration of the distribution of perturbed reach endpoints x_{fb} for reaches with aimpoint x_{aim} (indicated by the black vertical line). **B**, Experimenter-imposed rewards (black) and penalties (gray). The probability of hitting the target or penalty is indicated by the black and gray cross-hatching in **A**. **C**, Expected gain as a function of aimpoint. The black square indicates the expected gain for the example in **A**. The black circle corresponds to the optimal aimpoint x_{opt} resulting in maximum expected gain.

Figure 3. Optimal aimpoint x_{opt} as a function of overall variance. The vertical lines indicate the range of experienced overall variance in our experiments

across participants and epochs. Optimal aimpoint is approximately a linear function of variance.

Figure 4. Data for two subjects. **A**, Perturbation magnitude σ_{pert} across experimental sessions 1 (solid line) and 2 (dashed line). **B**, Data points: Reach endpoints plotted as shifts away from the penalty region relative to target center. Data are plotted only for penalty trials in which the reach was completed within the time constraint. Lines: Aimpoints based on the optimal omniscient strategy, i.e., assuming knowledge of both σ_{motor} as well as the current trial's value of σ_{pert} .

Figure 5. Compensation for uncertainty. **A**, Scatterplot of reach endpoint shifts away from the penalty region as a function of the optimal shift using the omniscient strategy. The solid identity line corresponds to the optimal omniscient strategy. The dashed regression line indicates the degree to which this participant responded to changes in outcome uncertainty. **B**, Regression slopes for all seven participants. Error bars are 95% confidence intervals. Six of seven participants had a significant degree of compensation for changing uncertainty. Most participants undercompensated, but one, SO, overcompensated.

Figure 6. Response to the previously experienced perturbation. A linear regression was computed to predict each trial's reach endpoint coded as a shift from the center of the target away from the penalty region as a function of the

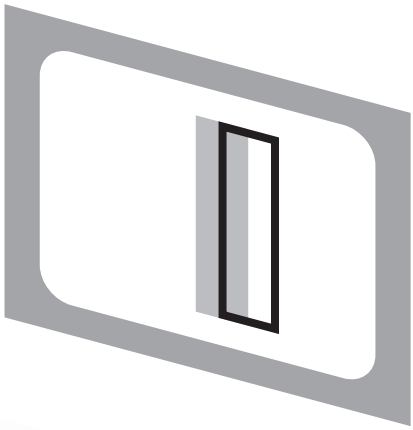
most recently experienced squared perturbation (**A** and **B**) or raw perturbation (**C** and **D**). Plotted are the regression weights (**A** and **C**, with 95% confidence intervals) and variance accounted for (**B** and **D**), as well as across-subjects averages and ± 2 SE.

Figure 7. Lagged regression. Results of a regression predicting each trial's reach endpoint coded as a shift from the center of the target away from the penalty as a linear combination of the previous 8 trials' squared perturbations. Gray: individual subjects. Black: Group average and ± 2 SE error bars. The isolated error bar at the upper right represents the average, over subjects and lags, of the 95% CI from the individual linear regressions.

Figure 8. Model comparison. Rectangles indicate each of the 7 models fit to the data and the number of model parameters for each (excluding the noise parameter). Arrows indicate model nesting, with the more complex model placed higher in the figure. Labels on the arrows indicate the results of nested hypothesis tests done individually for each participant.

Figure 9. Exponential model versus lagged regression. **A**, Lagged regressions (solid lines, from Fig. 7) and exponential weights (from fits of the "Exponential variance only" model) for each participant. Note the y-axis scale varies across subjects. **B**, Scatterplot of lagged regression vs. exponential model weights.

A



B

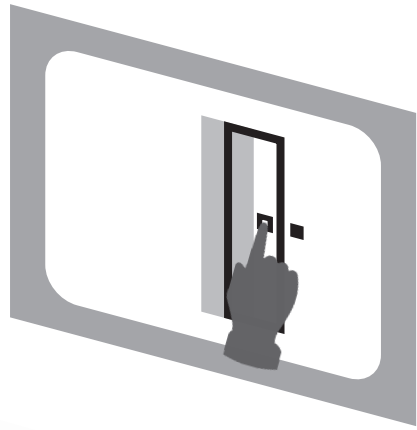


Figure 1

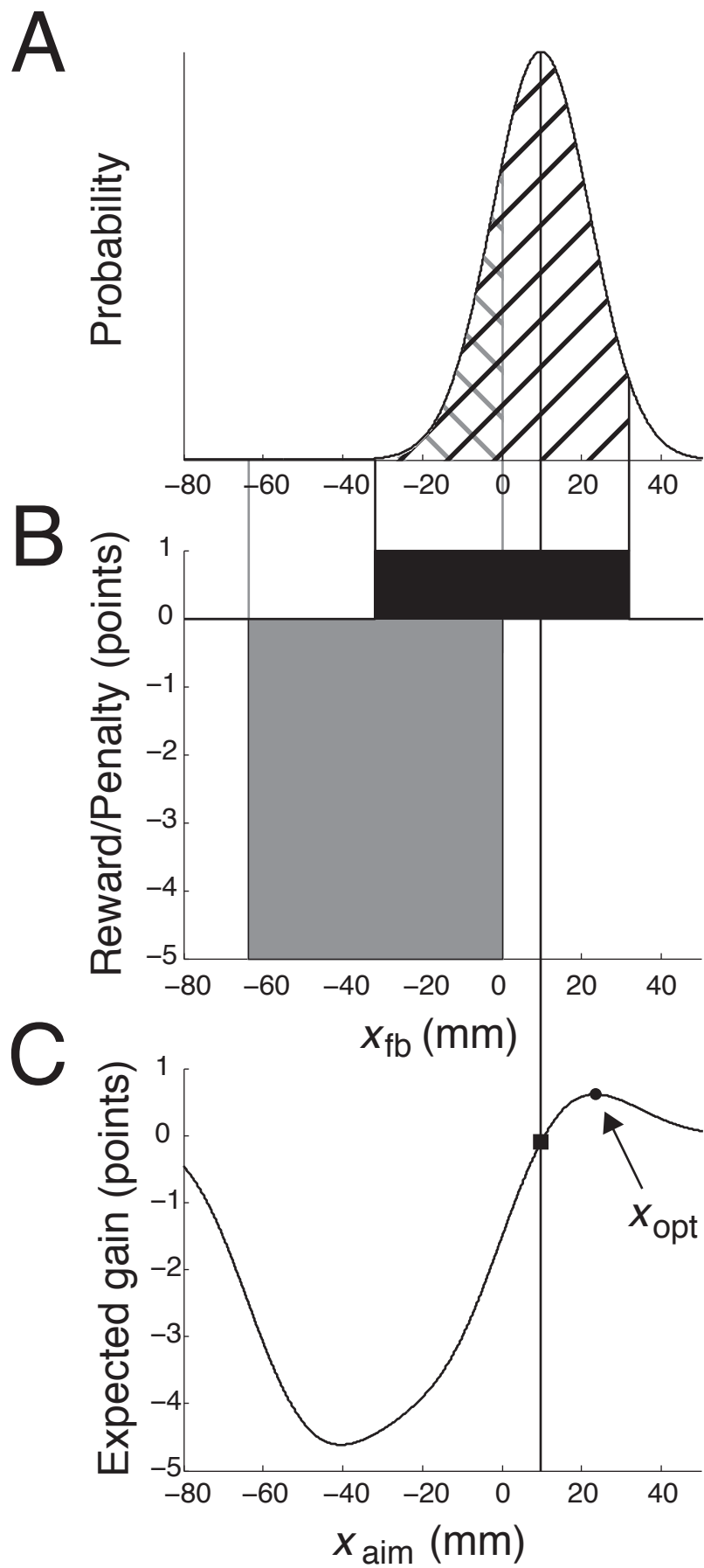


Figure 2

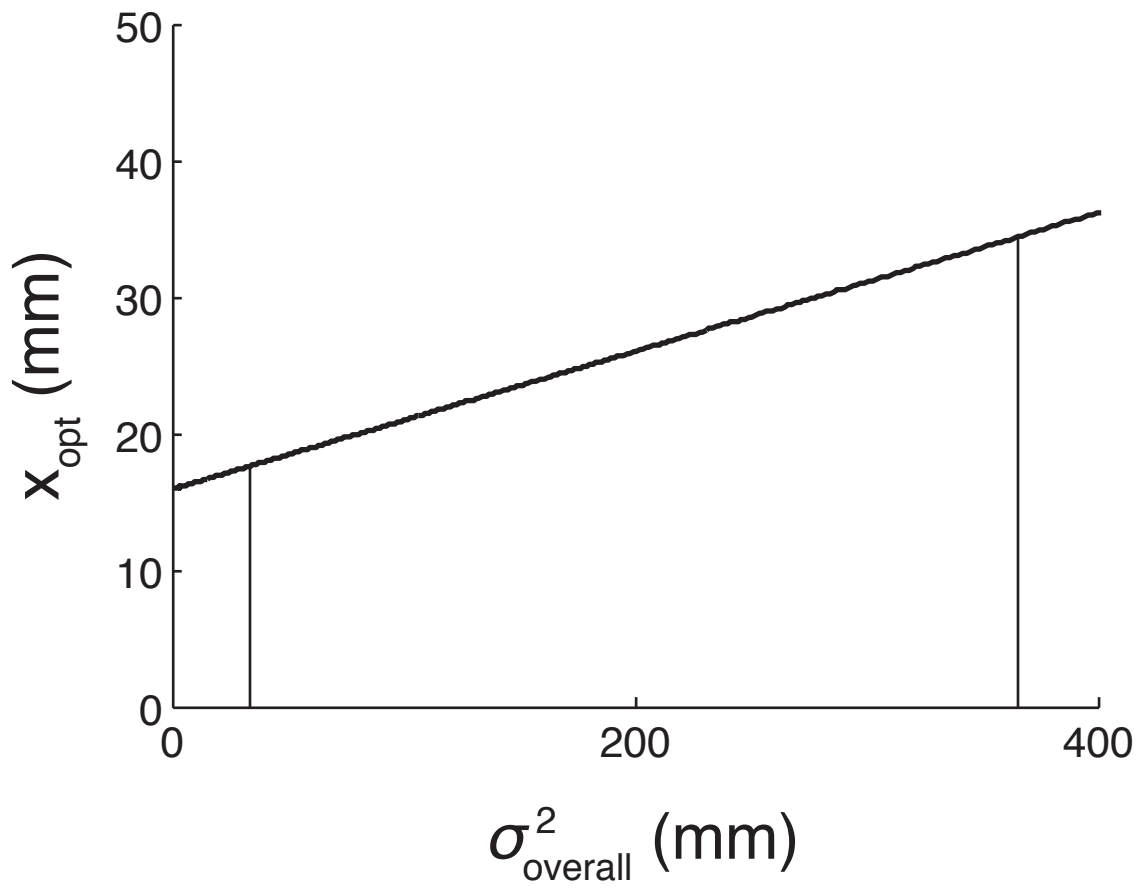


Figure 3

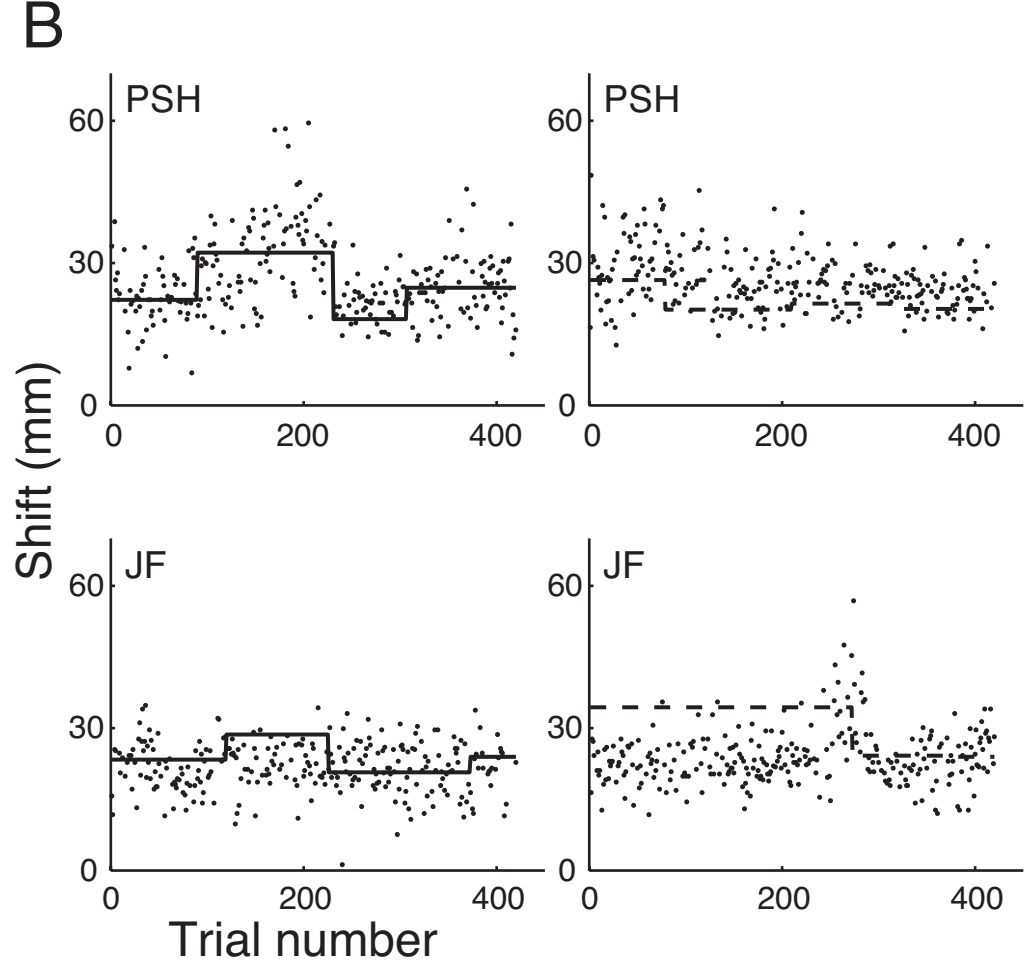
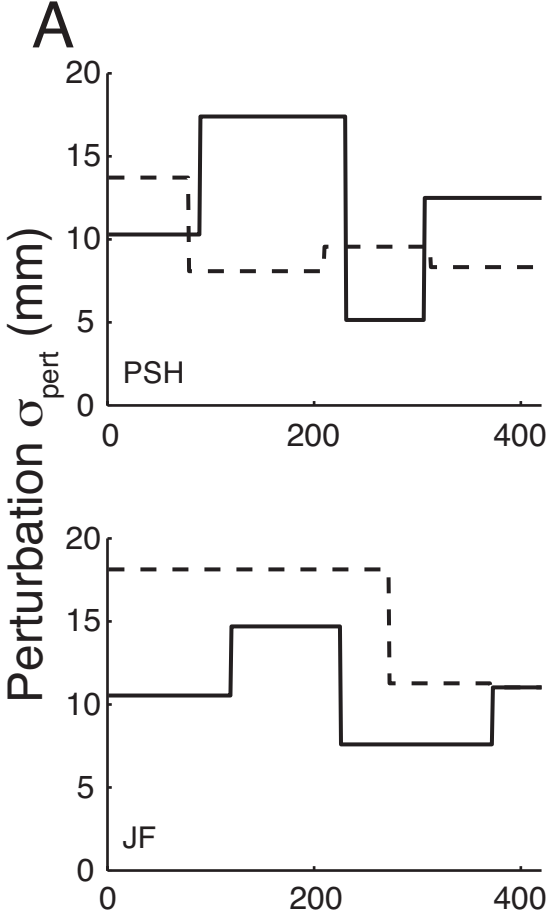


Figure 4

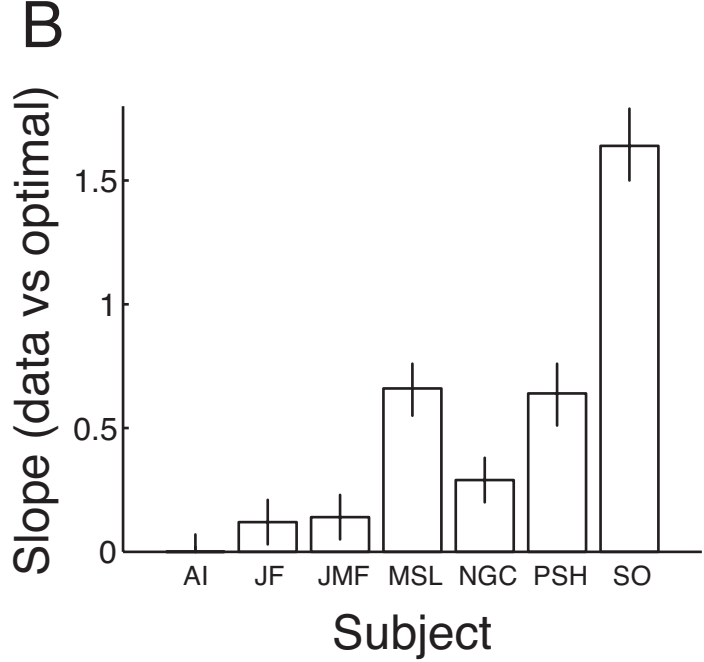
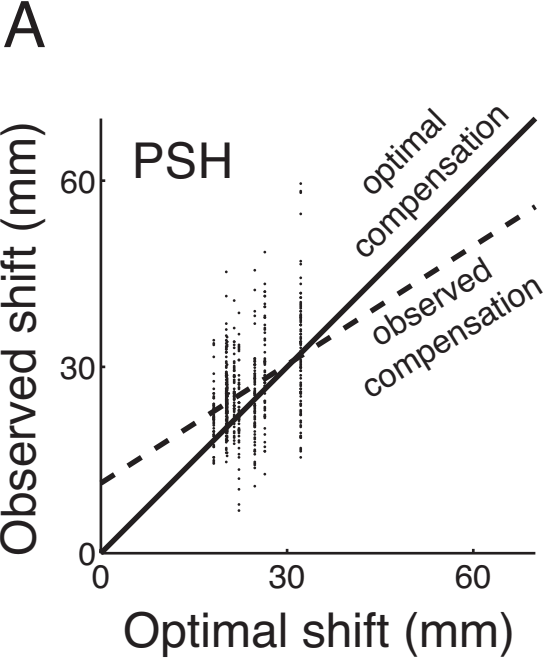
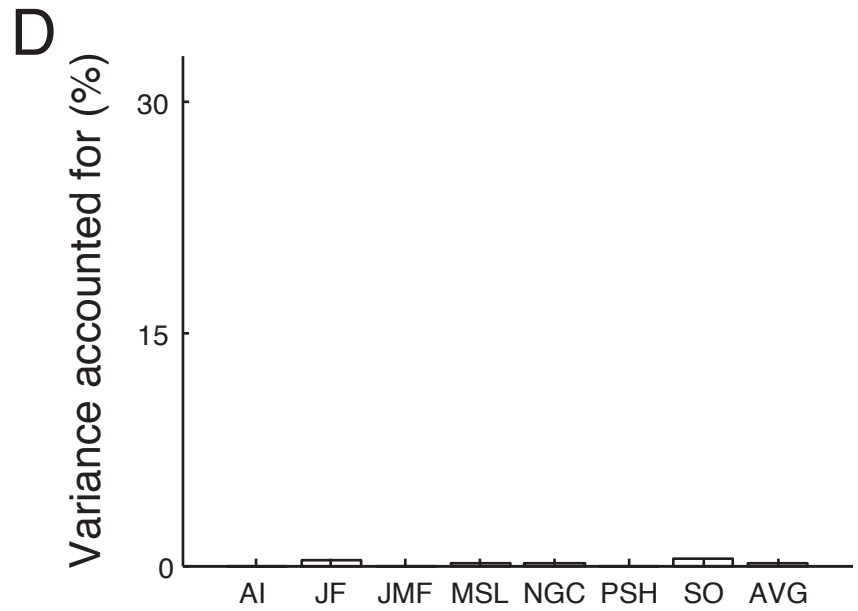
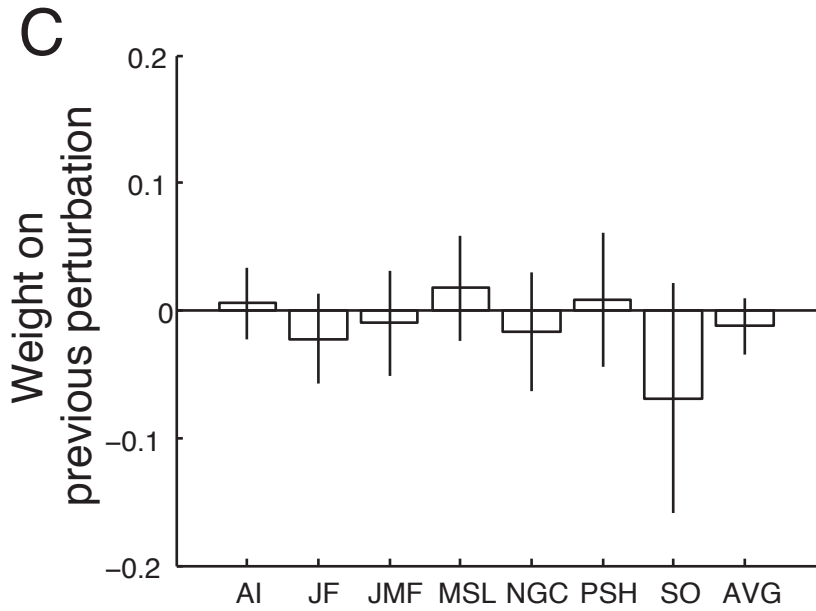
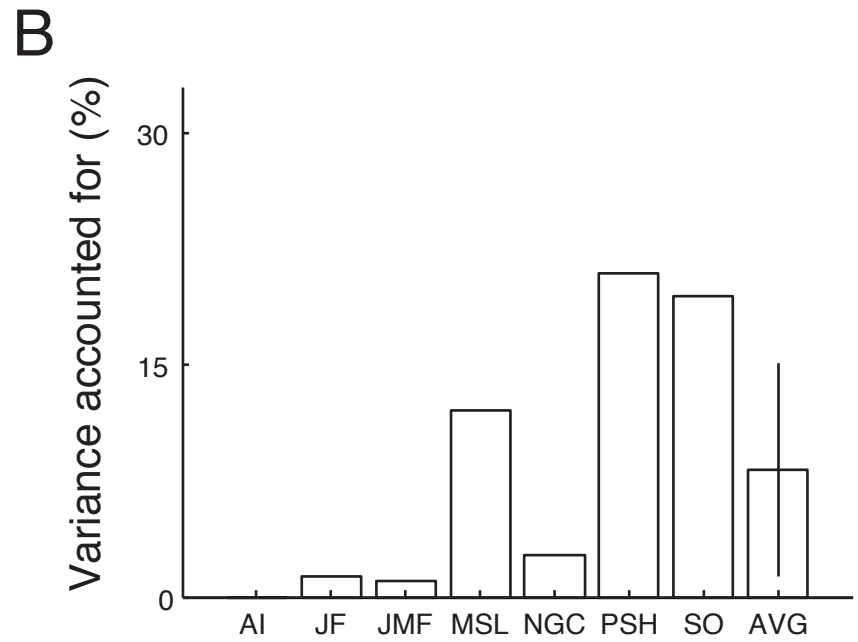
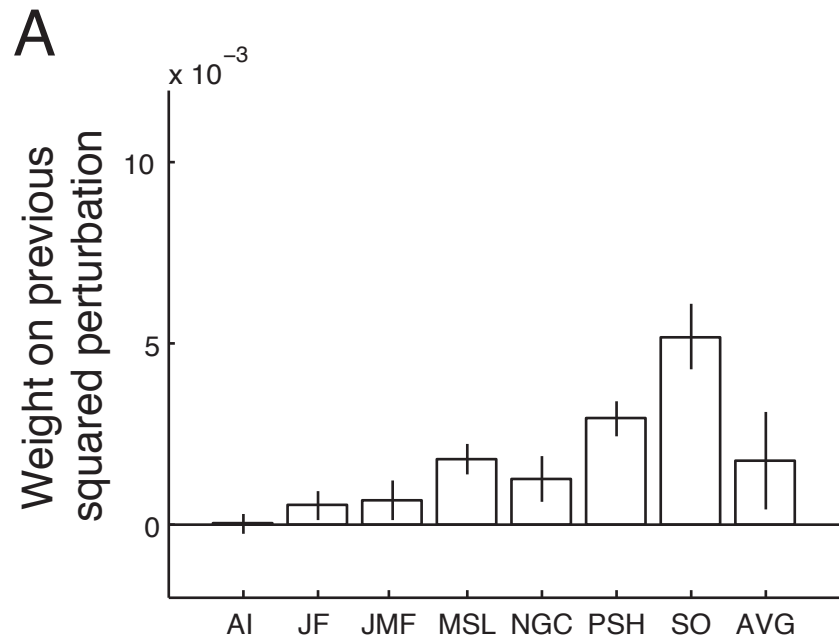


Figure 5



Subject

Figure 6

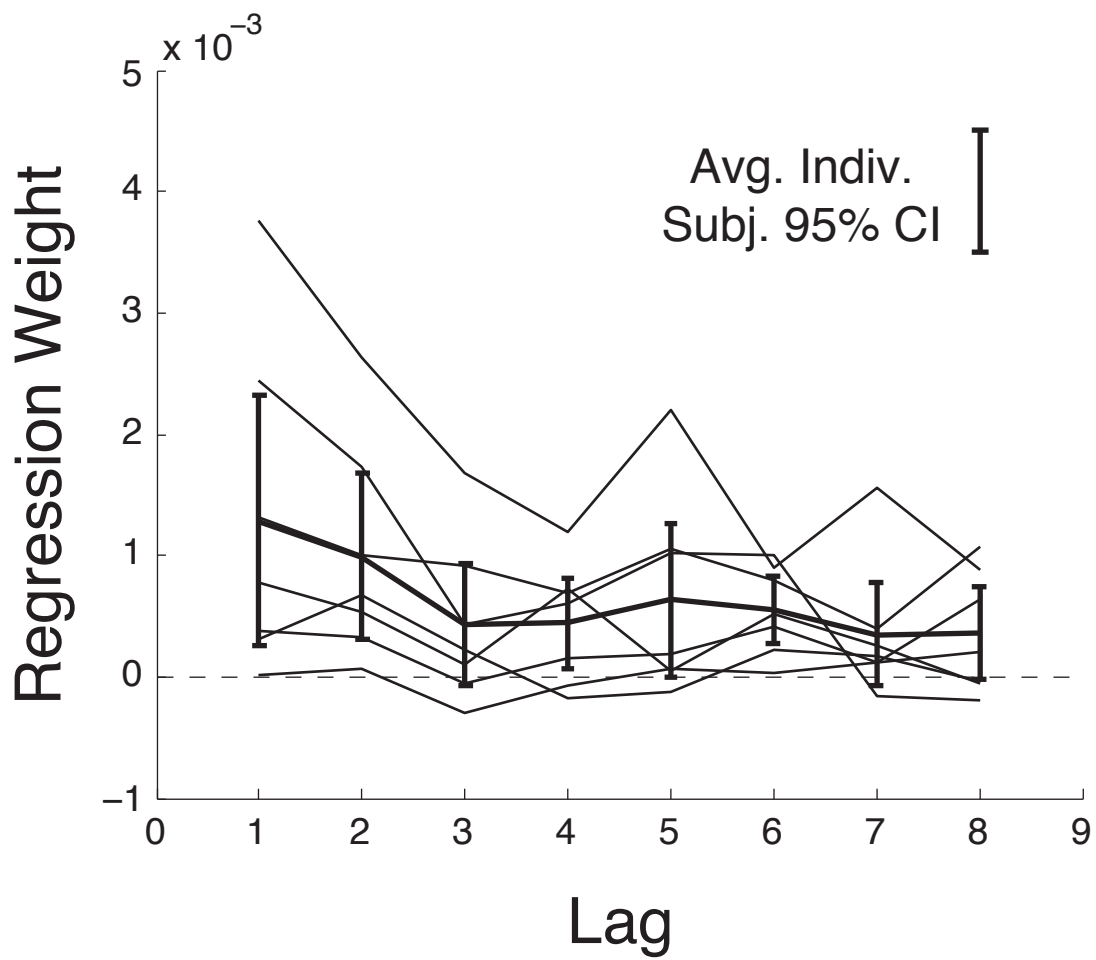


Figure 7

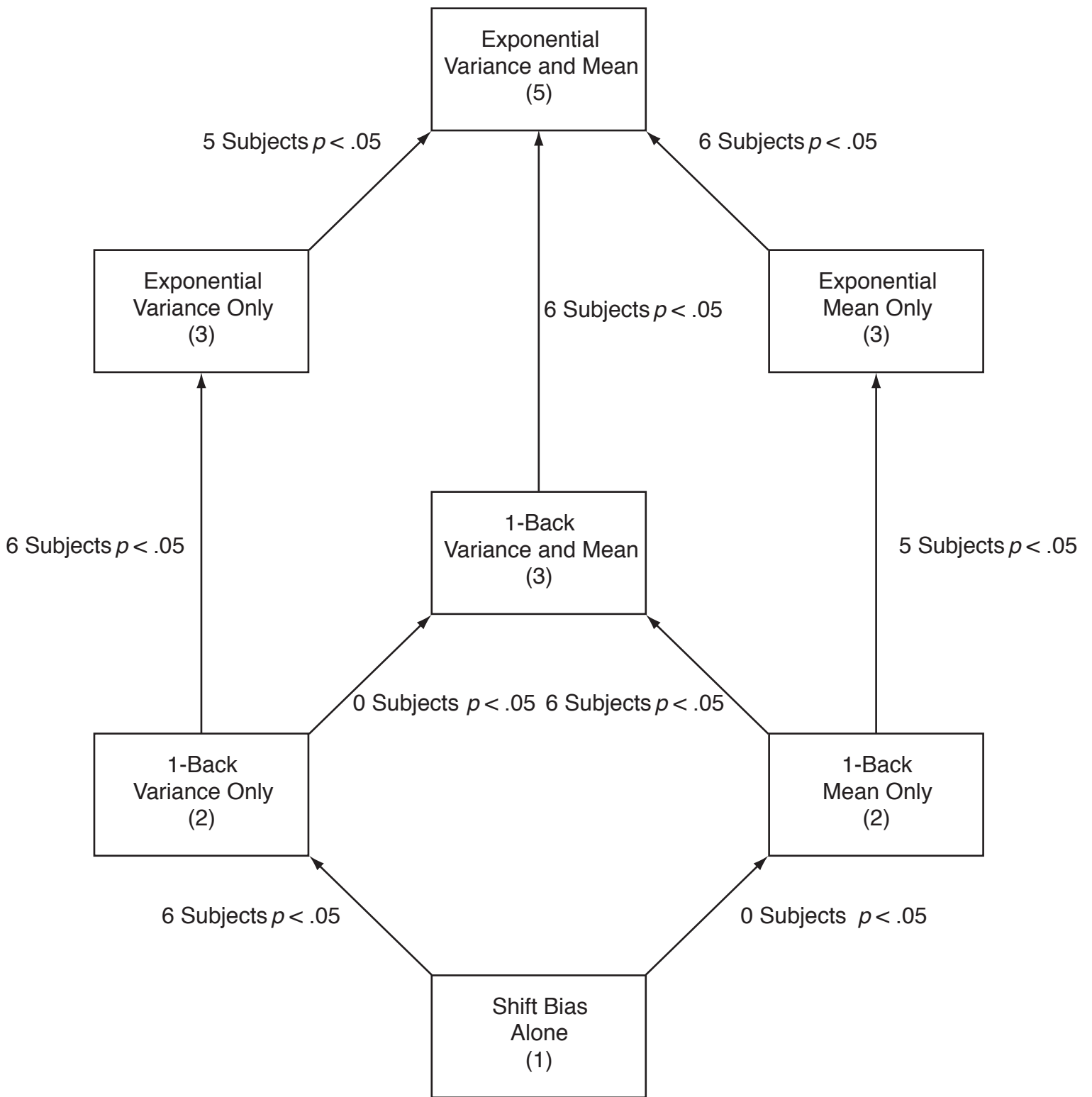


Figure 8

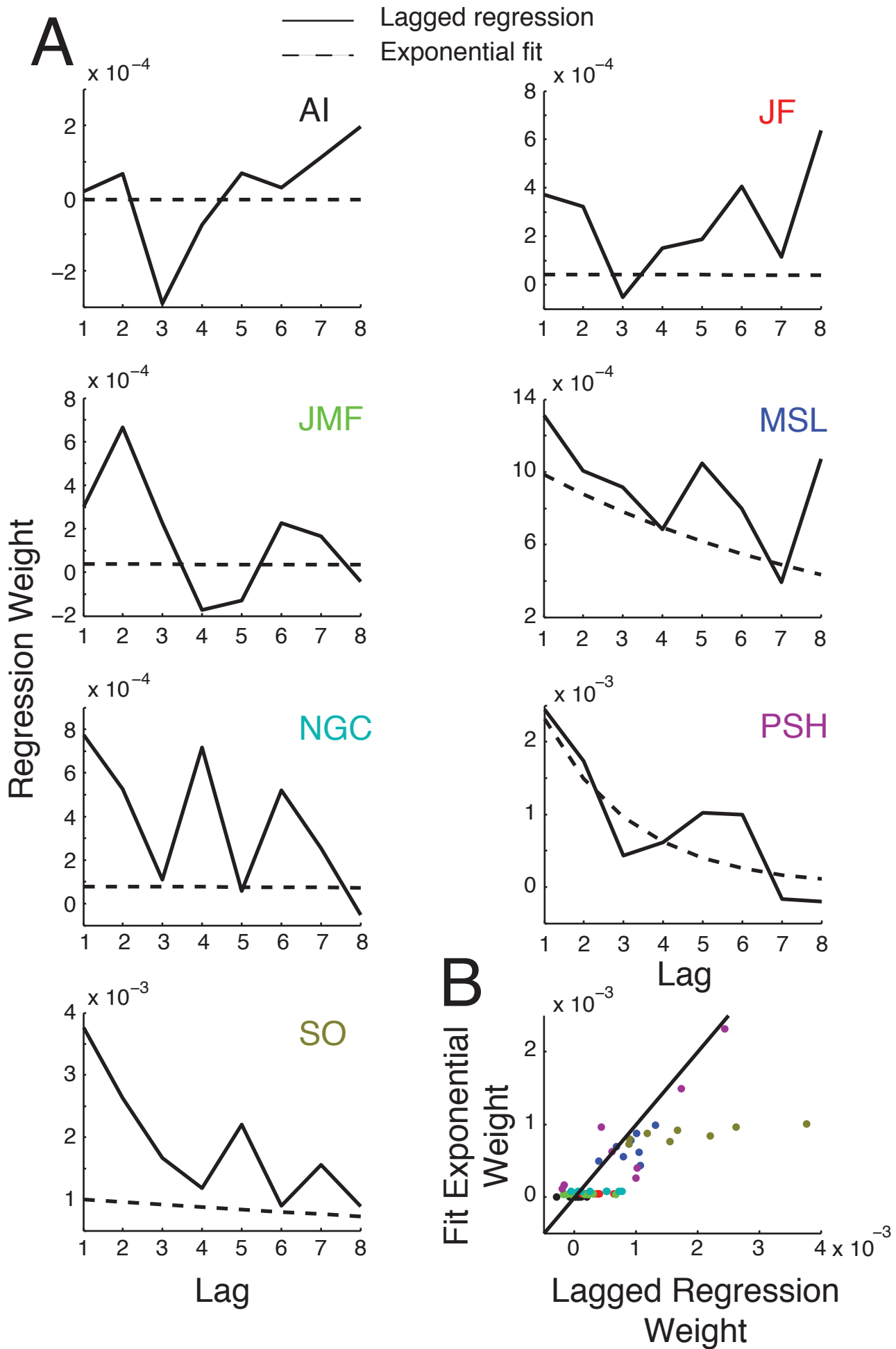
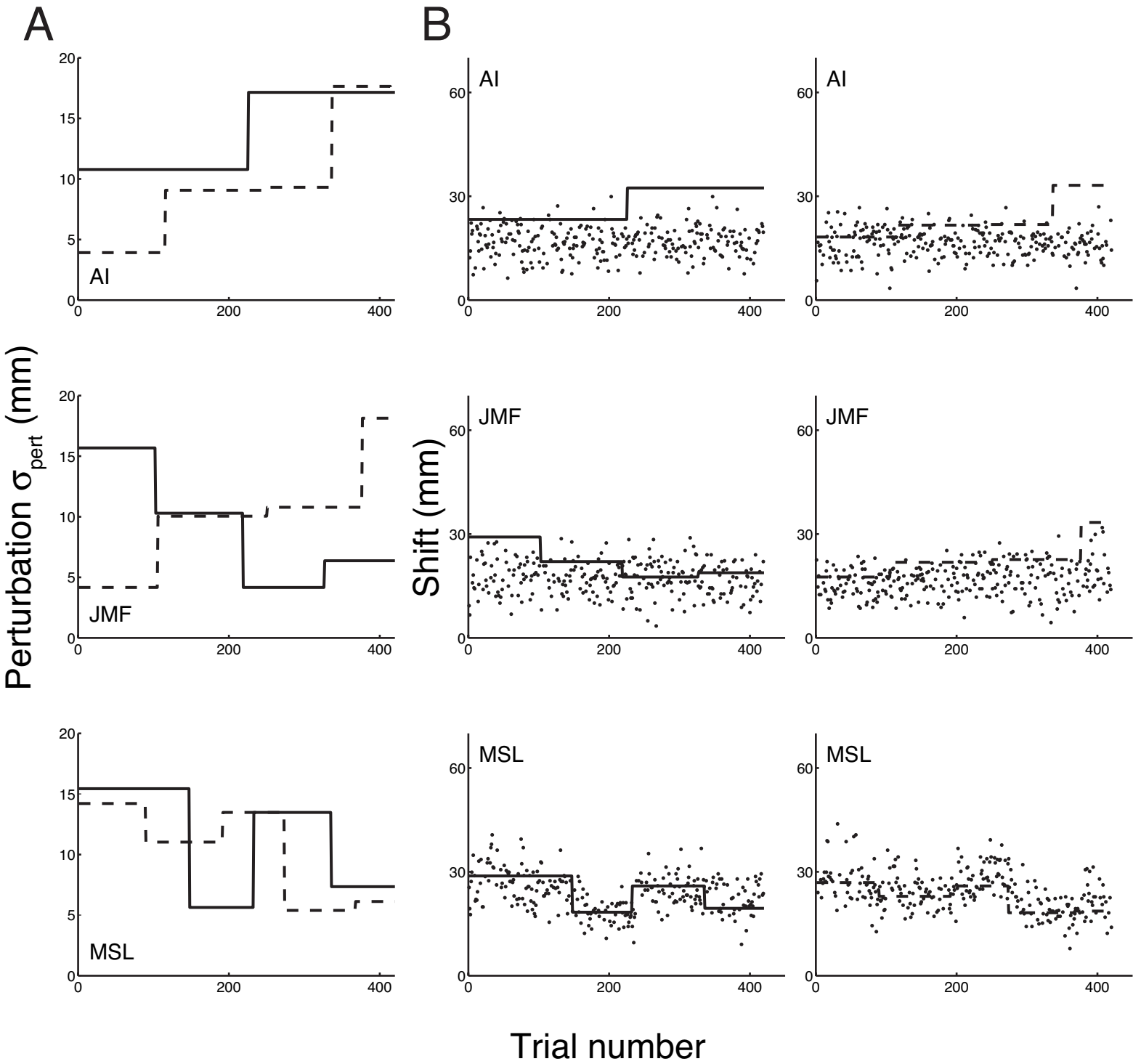


Figure 9



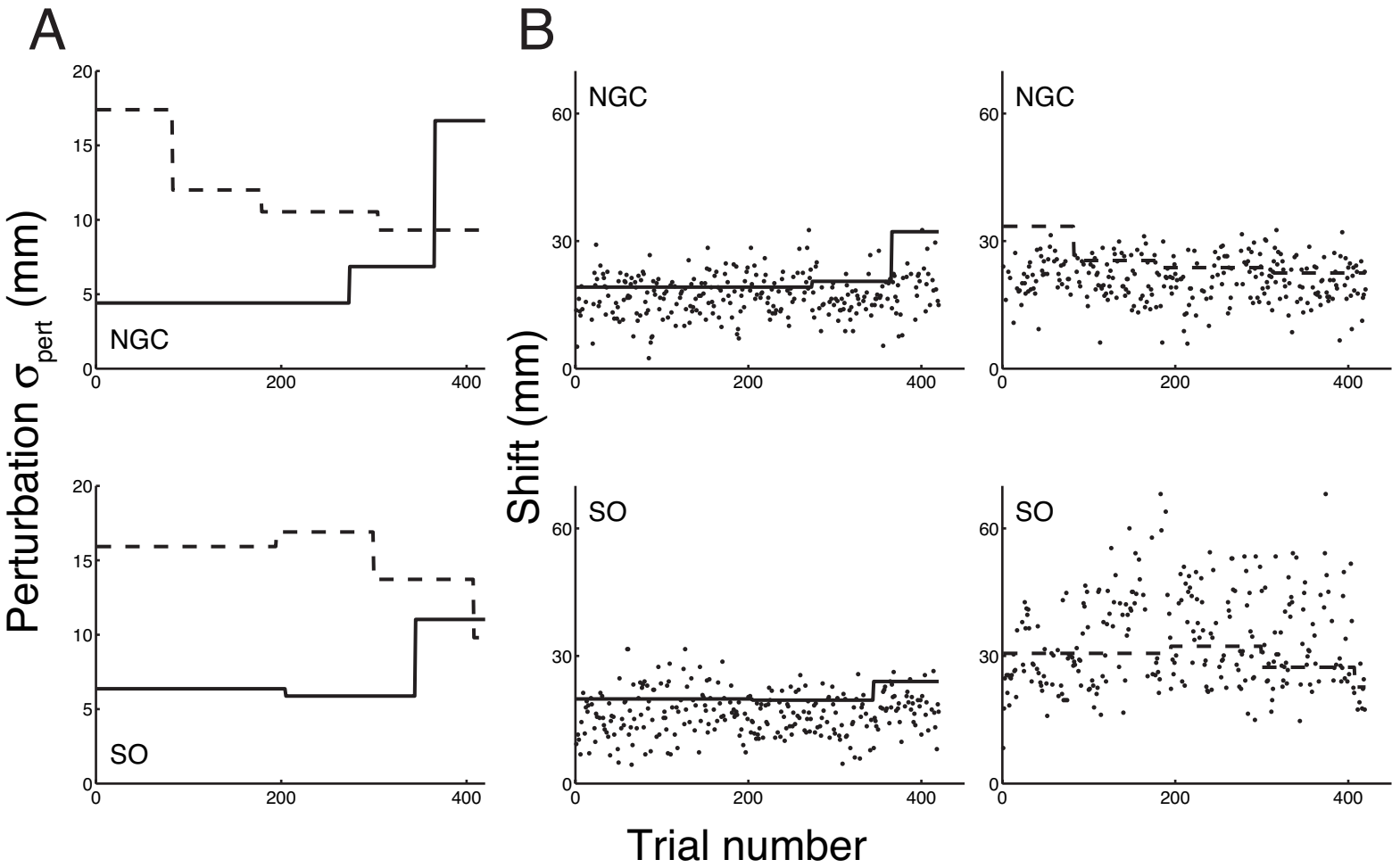


Figure S1 (contd.)