Supplementary Material: Strategies for exploration in the domain of losses

Paul M. Krueger[1,*], Robert C. Wilson[2,*], and Jonathan D. Cohen[3,4]

[1] Department of Psychology, University of California, Berkeley 94720
[2] Department of Psychology and Cognitive Science Program, University of Arizona 85721
[3] Princeton Neuroscience Institute, Princeton University 08544
[4] Department of Psychology, Princeton University 08544
[*] equal contribution

Full instructions for the task

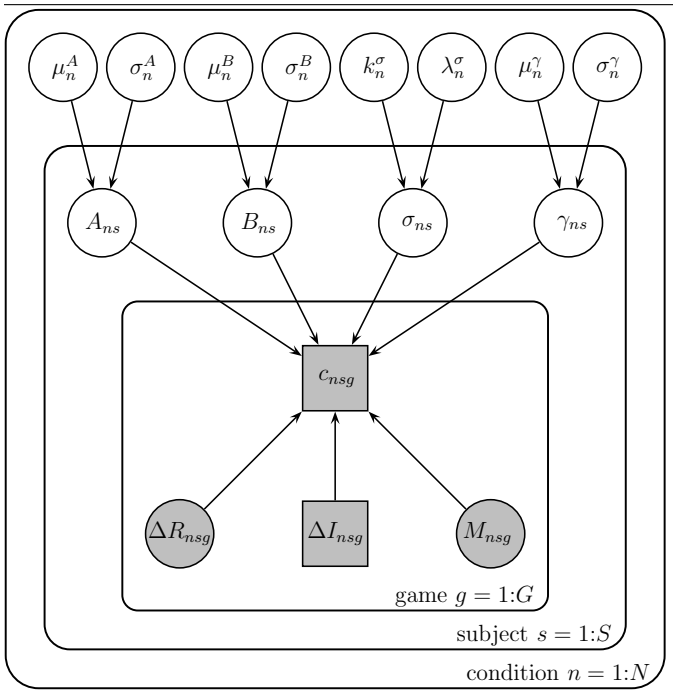
Before beginning the task, participants read a set of illustrated on-screen instructions. Each bullet point below shows text from a single screen (illustrations are omitted here to save space). The order in which participants were introduced to the gains and losses conditions, and all references to the tasks thereinafter, as well as the final example, reflected the block order of gains and losses for each particular participant. The example below is one in which the losses condition came first.

- Welcome! Thank you for participating in this experiment.
- In this experiment we would like you to choose between two one-armed bandits of the sort you might find in a casino.
- The one-armed bandits will be represented like this
- For the first half of the experiment, your task is to minimize how many points you lose overall. This is called the LOSSES task.
- For the LOSSES task, every time you choose to play a particular bandit, the lever will be pulled like this ...
- ... and the amount of points lost will be shown like this. For example, in this case, the left bandit has been played and is subtracting 23 points.
- For the second half of the experiment, your task is to maximize how many points you gain overall. This is called the GAINS task.

- The GAINS task is played similarly to the LOSSES task, but with points added to your overall payment ...
- For example, in this case, the left bandit has been played and is adding 77 points.
- The points you lose and gain by playing the bandits will be converted into REAL money at the end of the experiment. Therefore, the fewer points you lose and the more points you gain, the more money you will earn.
- A given bandit tends to subtract (in the LOSSES task) or add (in the GAINS task) the same amount of points on average, but there is variability in the amount on any given play.
- For example, if you're playing the LOSSES task, the average points subtracted for the bandit on the right might be 50, but on the first play we might see -48 points because of the variability ...
- ... on the second play we might see -44 points ...
- ... if we open a third box on the right we might see -55 points this time ...
- ... and so on, such that if we were to play the right bandit 10 times in a row we might see these points ...
- If you're playing the GAINS task, the average points added for the bandit on the right might be 50, but on the first play we might see 52 points because of the variability ...
- ... on the second play we might see 56 points ...
- ... if we open a third box on the right we might see 45 points this time ...
- ... and so on, such that if we were to play the right bandit 10 times in a row we might see these points ...
- Both bandits will have the same kind of variability and this variability will stay constant throughout the experiment.
- One of the bandits will always subtract fewer points (on the LOSSES task) or add more points (on the GAINS task) and hence be the better option to choose on average.
- When you move on to a new game, then the average amount of points of each bandit will change.
- To make your choice: Press $<$ to play the left bandit. Press $>$ to play the right bandit
- On any trial you can only play one bandit and the number of trials in each game is determined by the height of the bandits. For example, when the bandits are 10 boxes high, there are 10 trials in each game ...
- ... when the stacks are 5 boxes high there are only 5 trials in the game.
- The first 4 choices in each game are instructed trials where we will tell you which option to play. This will give you some experience with each option before you make your first choice.

- These instructed trials will be indicated by a green square inside the box we want you to open and you must press the button to choose this option in order to move on to see the outcome and move on the next trial. For example, if you are instructed to choose the left box on the first trial, you will see this:
- If you are instructed to choose the right box on the second trial, you will see this:
- Once these instructed trials are complete you will have a free choice between the two stacks that is indicated by two green squares inside the two boxes you are choosing between.
- The first half of the experiment will be the LOSSES task, so remember to try to minimize the overall number of points lost. You will be notified when you're halfway through the experiment, before the task changes.
- Press space when you are ready to begin. Good luck!

Reward magnitude model



Group-level parameters

$\mu_n^A \sim \text{Gaussian}(0, 1000)$
$\sigma_n^A \sim \text{Gamma}(1, 0.001)$
$\mu_n^B \sim \text{Gaussian}(0, 1000)$
$\sigma_n^B \sim \text{Gamma}(1, 0.001)$
$k_n^\sigma \sim \text{Exponential}(0.001)$
$\lambda_n^\sigma \sim \text{Exponential}(0.001)$
$\mu_n^\gamma \sim \text{Gaussian}(0, 1000)$
$\sigma_n^\gamma \sim \text{Gamma}(1, 0.001)$

Subject specific parameters

$A_{ns} \sim \text{Gaussian}(\mu_n^A, \sigma_n^A)$
$B_{ns} \sim \text{Gaussian}(\mu_n^B, \sigma_n^B)$
$\sigma_{ns} \sim \text{Gamma}(k_n^\sigma, \lambda_n^\sigma)$

Observed choices

$p_{nsg} \leftarrow \left[1 + \exp\left(\frac{\Delta R_{nsg} + A_{ns}\Delta I_{nsg} + B_{ns} + \Delta I_{nsg} M_{nsg}\gamma_{ns}}{\sigma_{ns}}\right)\right]^{-1}$

$c_{nsg} \sim \text{Bernoulli}\,(p_{nsg})$

Figure S1 – Graphical representation of the reward magnitude model.

Model of optimal behavior

Adapted from Wilson et al. (2014).

We modeled optimal behavior by solving a dynamic programming problem that computes the action that will produce the maximum expected outcome over the course of a game. The model knows that the mean outcomes are generated from a truncated Gaussian distribution with a given variance. It treats the gains and losses conditions equivalently. The optimal model solves a dynamic programming problem (Bellman, 1957; Duff, 2002) to compute the action that will maximize the expected total reward over the course of each game.

To do this the model first infers a distribution over the mean of each option given the observed rewards. We write r_t to denote the reward on trial t in the game, c_t to be the choice on trial t and D_t to be the set of choices and rewards up to and including time t. We assume that the model knows that the rewards are generated from a truncated Gaussian distribution and we further assume that it knows that the standard deviation of this distribution, σ_n.

In this case, the inferred distribution over the mean of option a, μ^a, given the history of choices and rewards is

$$p(\mu^a|D_t) \propto \sqrt{\frac{n_t^a}{2\pi}}\frac{1}{\sigma_n}\exp\left(-\frac{n_t^a(\mu^a - R_t^a/n_t^a)^2}{2\sigma_n^2}\right)p(\mu^a) \tag{1}$$

where n_t^a is the number of times option a has been played, R_t^a is the cumulative sum of the rewards obtained from playing option a and $p(\mu^a)$ is the prior of the mean. In our model we assumed an improper, uniform prior on μ^a (although we should note that it is straightforward to include a Gaussian prior instead). With this prior, equation (1) shows that the model's state of knowledge about option a is summarized by the two numbers, n_t^a and R_t^a. We can thus define the *hyperstate* (Duff, 2002), S_t, the state of information that the model has about both options as

$$S_t = (n_t^A, R_t^A, n_t^B, R_t^B). \tag{2}$$

With the hyperstates defined in this way we can now specify a Markov decision process within this state space. In particular we can define a transition matrix, $T(S_{t+1}|S_t, a)$, which describes the probability of transitioning between states S_{t+1} and S_t given action a. To compute this we note that if action $a = A$ is chosen on trial t and reward r_t is observed, then new state on the next trial will be

$$S_{t+1} = (n_t^A + 1, R_t^A + r_t, n_t^B, R_t^B). \tag{3}$$

Further, given the distribution over the mean, using equation (1) we can predict that this outcome will occur with probability

$$p(r_t|S_t, A) = \int d\mu^A p(r_t|\mu^A) p(\mu^A|S_t)$$

$$= \sqrt{\frac{n_t^A}{2\pi(1+n_t^A)}} \frac{1}{\sigma_n} \exp\left(-\frac{(r_t - R_t^A/n_t^A)^2}{2\sigma_n^2}\right) \tag{4}$$

Note that this result comes because both $p(r_t|\mu^a)$ and $p(\mu^a|D_t)$ are Gaussians, with $p(\mu^a|D_t)$ defined in equation (1) and

$$p(r_t|\mu^a) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(r_t - \mu^a)^2}{2\sigma_n^2}\right) \tag{5}$$

In practice, to make the algorithm tractable we only consider a subset of possible outcomes, focusing on a set of 51 possible outcomes between 0 and 100 for the horizon 1 case and 21 possible outcomes in the horizon 6 case. Given this approximation we can then compute the set of possible states encountered during the task and solve the dynamic program by iterating the equations for the state values

$$V(S_t) = \max_a Q(a, S_t) \tag{6}$$

and the action values

$$Q(a, S_t) = \sum_{S'_{t+1}} T(S_{t+1}|S_t, a)(r_t(S_{t+1}) + V(S_{t+1})) \tag{7}$$

In particular we start at the last trial, $t = H$, and work backwards in time to the first trial. Here, by definition the action value is just the expected value of the reward from each option; i.e.,

$$Q(a_H, S_H) = \frac{R_H^{a_H}}{n_H^{a_H}} \tag{8}$$

Finally the optimal action is to choose the option for which has the highest value on the first free trial, i.e.

$$c_1 = \arg\max_a \; Q(a, S_1) \tag{9}$$

This analysis allows us to compute the optimal behavior on the task. To compute the optimal performance shown in Figure 3, we simulated choices from this optimal model on the same set of problems faced by the participants. We then computed performance in the same way as we did for humans (see Methods).

Choice curves analysis

Focusing our analyses on the first free-choice trial, we computed p_a, the probability of choosing bandit a over bandit b, as a function of the difference in observed mean of each bandit, using Equation 2. The parameters in Equation 2 were set as the mean of the estimated posterior distribution across participants. In the [1 3] unequal certainty condition, bandit a was defined as the lesser known bandit (i.e. the bandit that had been observed only once during the forced trials); in the [2 2] equal certainty condition, bandit a was arbitrarily defined as the bandit on the right. The resulting choice curves are shown in Figure S2, along with empirical averages across participants. The error-bars on the empirical data points indicate the standard error of the mean across participants.

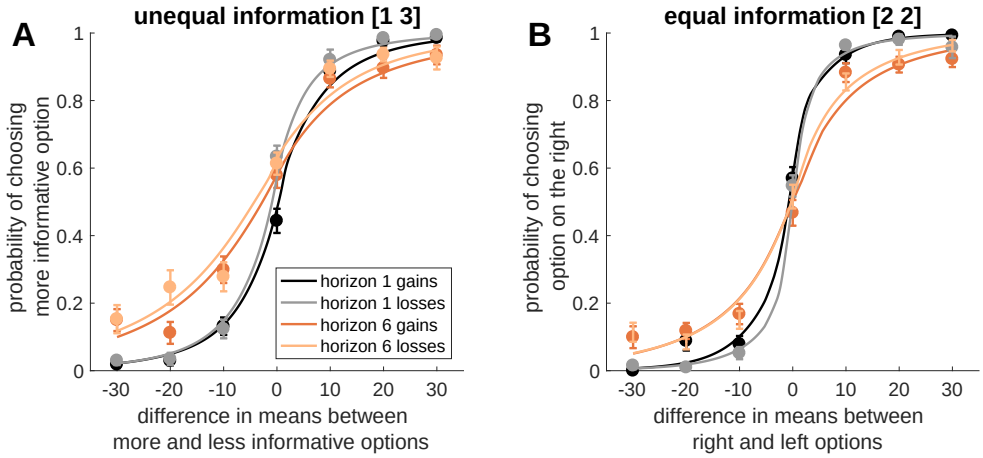


Figure S2. Choice curves for the first free-choice trial in the (A) [1 3] unequal and (B) [2 2] equal uncertainty conditions. Filled circles show experimental data averaged across participants, with error-bars indicating the standard error of the mean across participants. Curved lines show model-derived probability functions averaged across participants. (**A**) The fraction of times the more informative bandit is chosen, as a function of the difference in means between the more and less informative options. Compared to horizon 1 trials (gray-scale curves), horizon 6 trials (orange curves) show a greater information bonus, indicated by a shift in the indifference point (the point at which participants are equally likely to choose either option) further away from zero on the x-axis, as well as an increase in decision noise, indicated by a flattening of the slope of the curve. Within each horizon condition, the shift in indifference point is greater for the losses condition (light curves) than the gains condition (dark curves), indicating a greater uncertainty seeking in the losses condition. However, the slope of the curves within each horizon task is no different for the gains condition and the losses condition, indicating no change in decision noise. (**B**) In the equal uncertainty condition, there is less decision noise compared to the unequal uncertainty condition, as indicated by the steeper slopes of the curves within each horizon condition. There was no difference observed between the gains condition and the losses condition in the equal uncertainty condition. There is no information bonus in the equal uncertainty condition since both options have been sampled twice.

Participants choices were sensitive to the difference in mean between the two options, such that when the difference was large, participants were likely to choose the more rewarding (or less punishing) option, but as the difference became smaller, participants were more likely to choose either of the bandits.

In line with our previous findings for gains alone (Wilson et al., 2014), in the [1 3] unequal certainty condition there was a shift in the indifference point of the choice curves (the point at which participants were equally likely to choose either option) between horizon 1 and horizon 6. This was true for both the gains and losses conditions, and is consistent with directed exploration driven by an information bonus on the value of the lesser known option. That is, when participants had a longer time horizon in which to explore, they were biased towards the lesser known option, in hopes that acquiring more information about it would allow them to make more informed decisions later on, and hence improve their outcome overall.

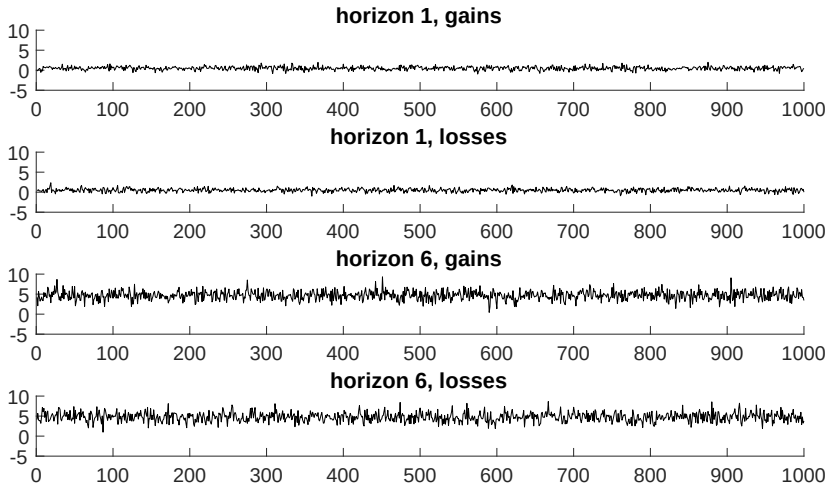
In addition to directed exploration, participants also showed random exploration, indicated by a flattening of the choice curve between horizons 1 and 6. This is also consistent with previous findings for gains (Wilson et al., 2014), and was equally true for both the gains and losses.

Comparing the gains and losses conditions, there was an overall increased bias toward the uncertain option for the losses condition, indicated by the overall leftwards shift in curves for the losses condition (light orange and grey curves), compared to the curves for the gains condition (dark orange and black curves; Figure S2A). Decision noise, indicated by the slope of the curve, does not change between gains and losses (Figure S2B).

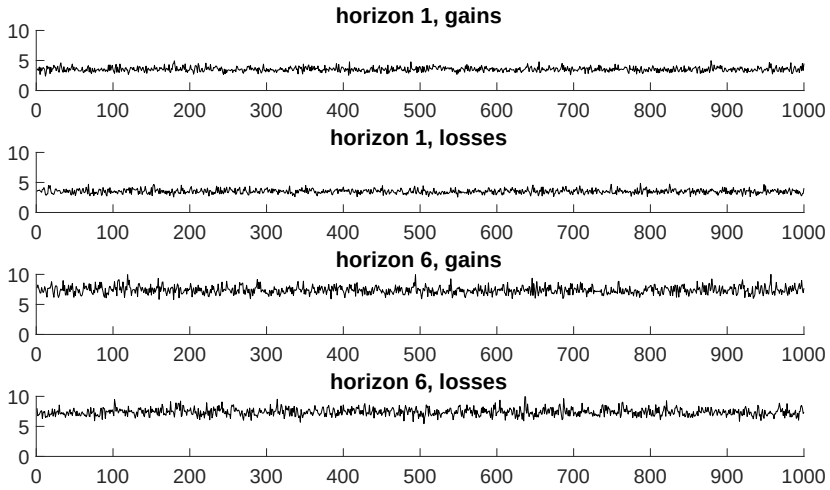
MCMC sampling convergence

As noted in the main text, all parameters were fit simultaneously using a Markov Chain Monte Carlo (MCMC) approach to sample from the joint posterior. We ran 4 separate Markov Chains with 500 burn-in steps to generate 1000 samples from each chain with a thin rate of 5. Below are serial plots of samples from one chain (after the burn-in) for the parameters shown in Figure 5: information bonus, [1 3] decision noise, and [2 2] decision noise.

Information bonus (μ^A):



[1 3] decision noise ($k^{\sigma}/\lambda^{\sigma}$):



[2 2] decision noise ($k^{\sigma}/\lambda^{\sigma}$):