Rewards are categories.

Erik J. Peterson
Dept. of Psychology
Colorado State University
Fort Collins, CO

Chapter 2 – Task and Models

On task

What they did, and when. The behavioral task each participant completed consisted of two parts. Depicted in Figure 1. (*top*), the first was a passive learning task wherein participants learned two rewarding perceptual categories by viewing randomly selected black and white sinusoidal gratings. Each grating (which was on-screen for 2 seconds) was followed by "Gain $1" or "Lose $1" in, respectively, green or red letters (1 second). The width of the grating's lines and their angles was derived from an information integration (category) distribution (Figure 2; borrowed from Spiering and Ashby (2008)). The disappearance of each grating and appearance of the reward was separated by an empty grey screen (1 second). Each trial terminated in a fixation cross (lasting at least 0.5 seconds). In total then, each trial lasted a total of 4.5 seconds. The trials for part 1 were spread over an initial training period completed outside the scanner, lasting 126 trials, and an in-scanner refresher lasting

45 trials. Prior to beginning training participants were, after some preliminaries, instructed to, "Attend to the screen in order to learn which types of gratings indicate wins and which types indicate losses". To minimize any stimulus specific effects, the category parameter distribution (Figure 2) to reward (i.e. "Gain $1" or "Lose $1") mapping was randomized for each participant.

Part 2 was a stimulus-response task that replaced classical rewards with an appropriate grating from task 1 (Figure 1, *bottom*). Gratings matching the Gain category were used for positive reinforcement, while gratings indicative of losses were used as negative reinforcers. Each trial began with an abstract black and white "tree" stimuli (left most image in bottom of Figure 1). Each "tree" determistically belonged to one of two response categories ("q" or "w"). Subjects indicated their response by button press using either the right or left index fingers on a magnet-compatible response box. The response window lasted up to 2.5 seconds, but ended as soon as a response was made. Immediately following response the "tree" was replaced with a blank grey screen, which was on-screen for half a second and was replaced with a feedback screen. If the response was correct a new, that is never before experienced, exemplar grating from the Gain distribution was used; if the participant was incorrect, a new Loss grating appeared instead. The use of novel gratings forced the subjects to classify each grating prior to infering its value. This necessary inference made these rewards incompatible with primary or secondary definitions. If no response was made, or the wrong button was pressed, the subject's reward was replaced with, "No response detected" (these trials were excluded from further analysis). Feedback always lasted for 1 second and was terminated by a

fixation cross (0.5 seconds).

For the instructions in part 2 participants were told, "Each tree belongs to either category q or w. Which is the correct answer though is random. The shape of trees is meaningless. To learn the correct response for each tree you must start by guessing. Use what you learned about the rewarding properties of the gratings from part 1 to learn the right responses. Remember, a random subset of the Gains and Losses are real. These mostly determine how much you'll earn for your participation. So try and earn as much money as possible". Instruction for both parts were given orally by the experimenter using a script and Figure 1 as a visual aid.

Over the course of part 2, participants learned to classify 6 "trees", randomly selected at the start of the experiment out of a pool of 22 possible. Each of the 6 were experienced a total of 28-32 times for a total of 199 trials. The order of the trials in the second half of part 1 and all of part 2 was determined using a genetic algorithm designed to optimize fMRI signal detection, among other considerations. Most relevant to behavioral analysis, trials were in pseudo-random order with second order counterbalancing. For remaining details see p**??**.

As part of fMRI data acquisition, 18 participants completed both parts of the task (10 female, mean age of 24, ranging from 21 to 32). Participants were compensated at a base rate of $15.00 earning up to $30 more depending on behavioral performance. For the last 30 trials the participant was be paid an additional dollar for every correct response and lost a dollar for every incorrect response. Before beginning experiment participants were told that some trials would count, but were not told

till after which trials (i.e. the last 30). The highest payout was $45, i.e.perfect performance, the lowest was 30, indicating near chance behavior for the last 30. The average payout was $40.23.

Of the 18 participants, two were removed from further analysis as they demonstrated inversed learning (Figure 3, see *107* and *110*). Despite reporting a full understanding of the reward contingencies from part 1, in part 2 these participants displayed significant and consistent decreases in performance through time. Had this learning been in the usual direction it would have been considered better than average performance. In post task interviews both reported feeling as if they performed above average. Once they were informed of their inverse performance neither believed it. It seems possible then that both correctly learned the perceptual characteristics but mis-mapped the value labels, i.e. they got "Gain" and "Loss" mixed up. Post-experiment interviews further suggested that both the discarded subjects were under high personal stress. One subject, who was a PhD student, had his competency exams the next day. The other completed a 60 mile bike ride an hour prior to participation. Combined these participants data suggest that the perceptual and verbal characteristics of the reward categories are independently accessible, and that the verbal label may be more labile than the perceptual distributions. However given the other participants consistent positive performance and rapid learning it seems these two were a curious but isolated anomaly (Figure 3).

Well Behaved Results. On an individual basis the lower confidence interval around the binomial fit of the the accuracy data rose to above the chance level (0.5)

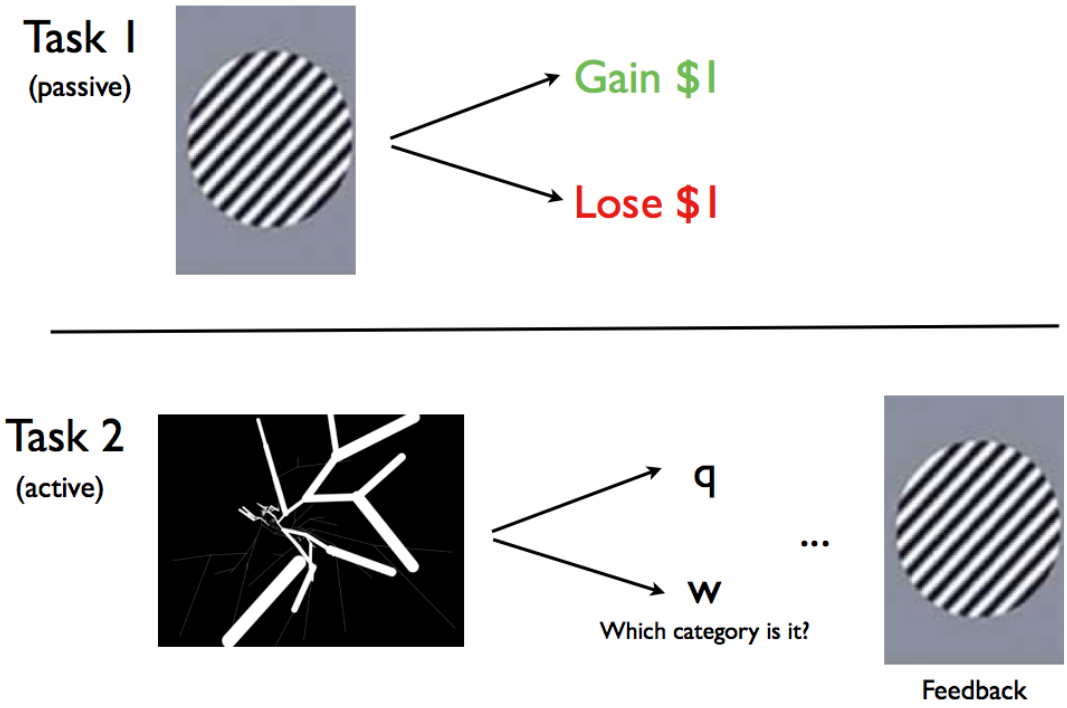

Figure 1. Depiction of the behavioral task. The *top* depicts part 1, the passive learning of the reward categories. The *bottom* depicts part 2, the stimulus-response learning phase.

by the last trial, except for participant 103, who did not learn (Figure 3). Many participants (11 of 16) greatly exceeded this minimum criterion, showing above chance learning by trial 10, and nearly all (14 of 16) exceeded chance by trial 20 (Figure 3). These individually good performances are reflected in the participants' aggregate performance, which was well above chance by trial 5 (Figure 5). This aggregate learning rate is consistent with past work in the lab using verbal or monetary feedback, indicating that the rewarding categories in present study are behaviorally similar to

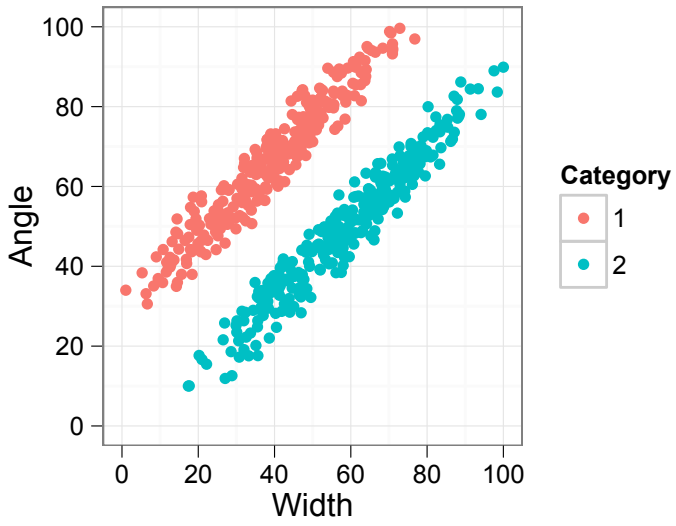

Figure 2. The two sinusoidal grating distributions for the information integration (II) category distributions. As II categories span the diagonal of the gratings parameter space (line width and angle successful learning requires consideration of both dimensions preventing participants from solving the categorization problem with simple rule-based strategies.

.

classical rewards. The consistency between classical rewards and this task were also reflected in the reaction time measures, which showed a 200 ms decrease over time, bottoming out near 850 (for individual averages see Figure 4 and for overall performance see Figure 6). Responses in similar, classically rewarding tasks end with reaction times near 700-800 milliseconds and show similar rates of decline. The 50-100 ms possible difference in reaction times may be due to the increased difficultly of classifying the rewards compared to simply reading the value of the outcome (e.g. "Gain $1").

3 Models and 2 Codes

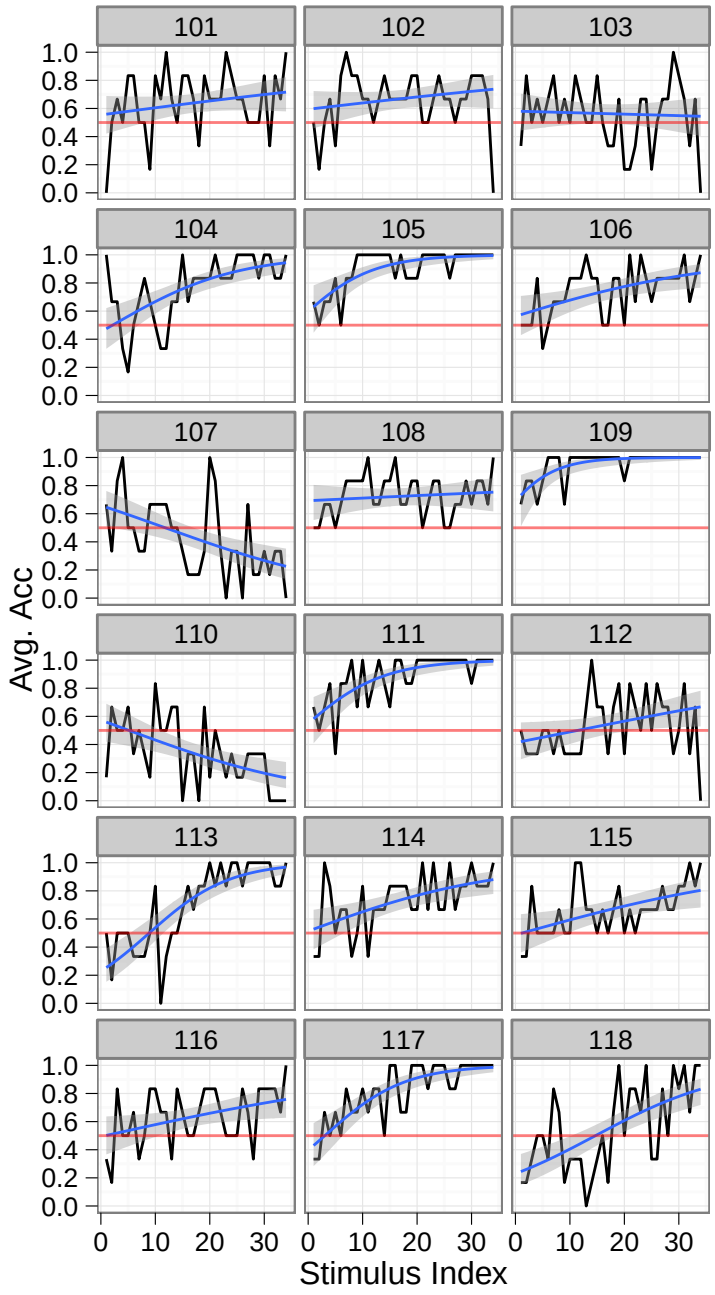

Figure 3. Average accuracy for each participant (black), averaged for all 6 stimuli by trial (i.e. Stimulus Index), blue line and the grey area represent a binomial regression fit of the data and bootstrapped 95% confidence intervals, respectively.

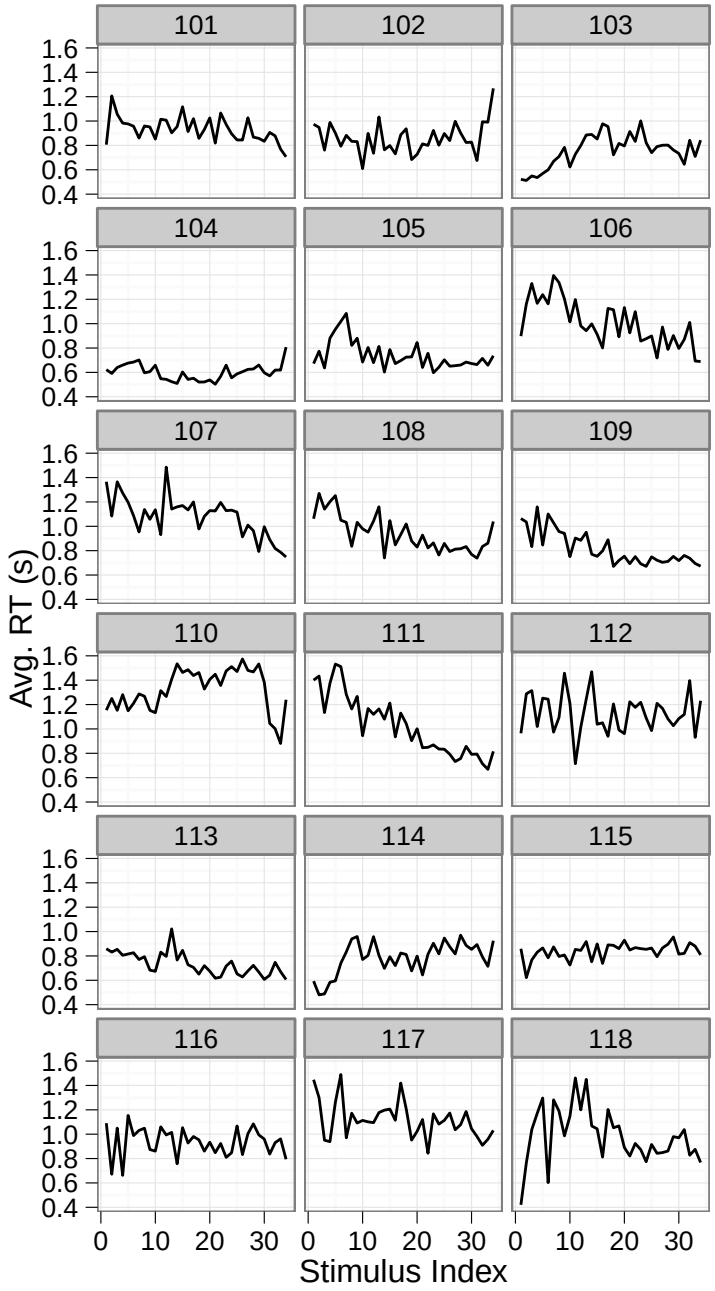

Figure 4. Mean reaction time for each participant, averaged for all 6 stimuli by trial (i.e. Stimulus Index).

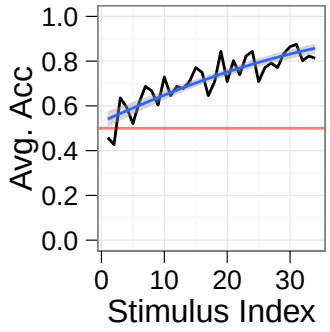

Figure 5. Mean accuracy (black), averaged for all 6 stimuli by trial, blue line and grey represent a binomial regression fit of the data and bootstrapped 95% confidence intervals, respectively.

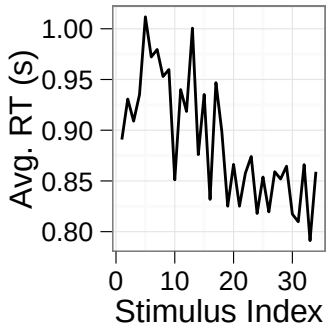

Figure 6. Mean reaction time (black), averaged for all 6 stimuli by trial, blue line and grey represent a linear regression fit of the data and bootstrapped 95% confidence intervals, respectively. See Fig 3 for learning criterion and other relevant details.

Our Three Models. Three Rescorla-Wagner models were constructed, each using a distinct reward representation. The first model treated rewards identically to a classical reward (e.g. a 0 for a loss or 1 for a gain). The second and third devalued each reward based on how similar it was to the category mean. This similarity metric is (necessarily) simplistic. As I reviewed on p**??**, there are many proposed models for how the categories are represented. Likewise, our understanding of the computaional implementation of the category learning systems is only just underway (Ashby & OBrien, 2005; Ashby & Ennis, 2006). As such there is no obvious way to interlace my hypothesis of rewarding categories with the category learning systems or with the many models of categorization they rely on. Avoiding such ambiguity, I took a simpler route driven by Shepard's basic finding (Shepard, 1987). Whatever the representation and/or category learning system that (may) implement rewarding categories, Shepard's work insists they show an exponential or Gaussian decline with similarity. For simple stimuli like a light or tone, similarity is measured from the initial training prototype (Guttman, 1956). However as my task does not have a singular prototype the mean of the parameters for all training trials (i.e. part 1) was used in its place. This simple substitution makes my categories identical to the simplest of the prototype category representations (Rosch, 1973; Ashby & Alfonso-Reese, 1995), making this a parsimonious yet literature driven first attempt to quantify similarity of rewards. Therefore, for model two similarity decreased exponentially measured from the training mean. While for model three it decreased along a normal Gaussian (for complete mathematical detail see p10).

Codes and Fits. For each participant and model, the two free parameters (α, which controls each model's rate of learning and β, which controls the steepness of the action selection criterion) were fit using an exhaustive[1] maximum log-likelihood search. Additionally each model was run using two separate reward coding schemes. In the first scheme Gains were valued as 1 and losses as 0. The second, which was based on the bivalent monetary value of the rewards, uses 1 and -1 for gains and losses respectively. The first scheme is universally used in human and animal modeling studies as well as in machine learning, and was Sutton and Barto (1998), recommended scheme. However recent recordings of dopaminergic neurons in monkey suggest that the true reward codes are complex, even perhaps redundant Kim, Shimojo, and O'Doherty (2006); Matsumoto and Hikosaka (2009). Among other schemes, they reported firing consistent with a bivalent reward code. Using the second scheme is mearly a first, albeit simple, step in incorperating the potential complexity of dopaminergic firing as observable by the fMRI BOLD signal and in human subjects.

TODO: discuss how the RW model is simplified and why.

The Incantations. To restate more formally, the Rescorla-Wagner model's value updates are defined by,

$$V(s,t) \leftarrow V(s,t) + \alpha * \delta \tag{1}$$

$$\delta = r_{classic}(t) - V(s,t) \tag{2}$$

[1] With a 0.05 precision, ranging from 0-1 for α and 0-5 for β

where $r_{classic}(t)$, i.e. the numerical representation of the rewards, can be coded as either

$$r_{classic}(t) = \{1, 0\} \tag{3}$$

or

$$r_{classic}(t) = \{1, -1\} \tag{4}$$

but where $r_{classic}(t)$ may also be replaced with

$$r_{exp}(t) = r_{classic}(t) * S_{exp} \tag{5}$$

or

$$r_{gauss}(t) = r_{classic}(t) * S_{gauss} \tag{6}$$

where the Euclidean distance

$$D = \sqrt{(\bar{\theta} - \theta)^2 + (\bar{W} - w)^2} \tag{7}$$

is transformed to a Shepard-like similarity metric (Shepard, 1987).

$$S_{exp} = e^{-D} \tag{8}$$

$$S_{gauss} = e^{-D^2} \tag{9}$$

Consistent with past work, all values are initialized at 0 (Beierholm, Anen, Quartz, & Bossaerts, 2011; Bischoff-Grethe, Hazeltine, Bergren, Ivry, & Grafton, 2009; Ger-

shman, Pesaran, & Daw, 2009)

$$V_{initial}(s,t) = 0. \tag{10}$$

and values are transformed to response selection probabilities via the softmax distribution (Sutton & Barto, 1998; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003).

$$p(s_1) = \frac{e^{\beta V(s_1,t)}}{e^{\beta V(s_1,a)} + e^{\beta V(s_2,a)}};\ s_1 = (s_i, q),\ s_2 = (s_i, w). \tag{11}$$

Fits and plots. On average none of the three models fit the accuracy data better than the rest (Figure 7). For brevity's sake each model will be referred to as "none", "exp" and "gauss", corresponding to Eq 2, Eq.5, and Eq. 6 respectively. Nor did the coding scheme impact the fits ("acc" and "gl", matching respectively Eq. 3 and Eq. 4; Figure 7)). For "acc" the step size parameter ("alpha" in Figure 8 matching α in Eq. 1) increased in "exp" compared to the other models. This increase was expected as the exponential similarity metric can sharply decrease the magnitude of each value update, requiring an increase in α to compensate. A similar trend was observed for the temperature parameter ("beta" in Figure 9, matching β in Eq. 11). The more equiprobable each action is the larger the temperature parameter. As such the increase for "exp" means participant's choices are more likely to change from trial to trial, which is again consistent with a decrease in update magnitudes.

Importantly the intra-subject variability in α and β was low, as demonstrated by the small standard error of both parameters (Figure 8 and 9). Consistent pa-

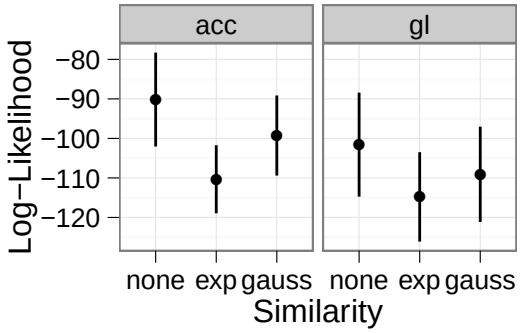

Figure 7. Average negative log-likelihoods for each of the models and coding schemes. Error bars represent standard errors.

rameter estimates between subjects support my use of subject-level parameters in the fMRI analyses, which in other hands have been reported to be too noisy to be reliable (N. D. Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Seymour, Daw, Dayan, Singer, & Dolan, 2007; O'Doherty et al., 2003). I believe, using subject-level parameters is a step crucial in assessing and maximizing model quality. The goal of any model of human behavior is to make good predictions for individual cases not just for aggregates of tens or hundreds of participants (N. Daw & Courville, 2007). However aggregates prediction is the norm in reinforcement learning models of human behavior (for examples see, N. D. Daw et al. (2011); Seymour et al. (2007); O'Doherty et al. (2003)).

The fit reinforcement learning models for every participant and coding scheme can be found in Figure 10 - 13. In the traditional formulation, where similarity does not impact reward value (e.g. see Figure 10 or 11) the reward prediction error decreases with learning, eventually plateauing at 0, for strong examples see subjects

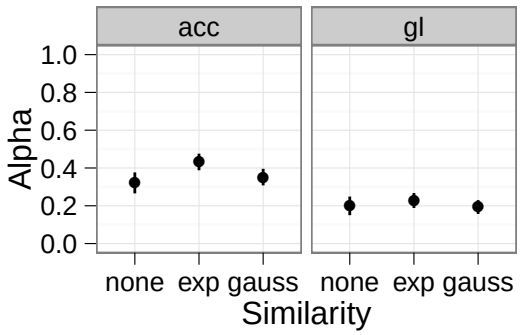

Figure 8. Average alpha values for each of the models and coding schemes. Error bars represent standard errors.

102, *105* and *111* in the "none" column of Figure 10. In contrast, both similarity models, "exp" and "gauss", never fully plateaued (again see Figure 10). In the context of the models, this is expected. Each grating will, in all likelihood, be non-identical to the mean. However as the parameter mean is the asymptotic expectation, reward prediction errors happen even after learning is complete. In the big picture, this is desired model behavior. The similarity adjusted model's are trying to capture the case where a reward's value may vary in ways that are not predictable *a priori* given the massive multiplicity of possible outcomes it is unlikely to find the same one multiple times. Small prediction error should then continue without end, as happens in these models.

When comparing "exp" and "gauss" models, you'll see the former appears to have lower magnitudes (Figure 10). Examining density plots composed of all participants data confirms this observation (Figure 14). The density plot also reveals that "exp" prediction errors are diminished more rapidly than their "gauss" counterparts

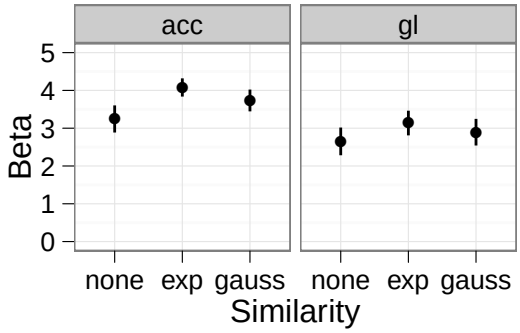

Figure 9. Average beta values for each of the models and coding schemes. Error bars represent standard errors.

(a pattern most clearly scene in the left panel of Figure 14).

Regardless of model class, between the two reward codes there are substantial differences in reward prediction behavior. The $\{1, -1\}$ (denoted in these plots as "gl") scheme leads to substantially more negative deflections that the $\{1, 0\}$ scheme ("acc") (compare Figure 11 and 10 as well as the *left* and *right* panels of 14).

ealues estimates for the two similarity adjusted rewards (i.e. "exp" and "gauss") were generally less than the alternative classic model ("none" in Figure 12 and 13). However, as consequence of their reduced dynamic range, the similarity adjusted model's value terms grew more rapidly, for example examine participants *105* and *109* in Figure 12. In these cases both the similarity value terms approached their maximum by trial 50 whereas "none" (the unadjusted term) took until trial 150. That is, taking into account the uncertainty of each reward's worth lead to an increase in the learning rate. This increase was independent of the reward coding scheme (i.e. see also Figure 12- 13).

The coding scheme's impact on the value terms was two fold. First, as losses lead to larger prediction errors using the "gl" scheme (-1 compared to 0), value increased faster for "acc" (see Figure 12 and 13 as well as 15). Second, and most strikingly, "gl" lead to negative value estimates of the undesirable choices (Figure 13 compared to 12). While, so far as I'm aware, no reinforcement learning model of human or animal have considered negative values estimates, there is empirical support. As reviewed on p**??**, orbital frontal and ventral medial frontal cortices encode the absolute value of rewarding or punishing outcomes (O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001; Hornak et al., 2004). As such, neural correlates of reinforcement learning derived negative value estimates might serve as an important link between theoretical and empirical findings on economic valuation. It might also serve as a link between reinforcement learning and affective/motivational processing (Knutson, Taylor, Kaufman, Peterson, & Glover, 2005; Delgado, Stenger, & Fiez, 2004).

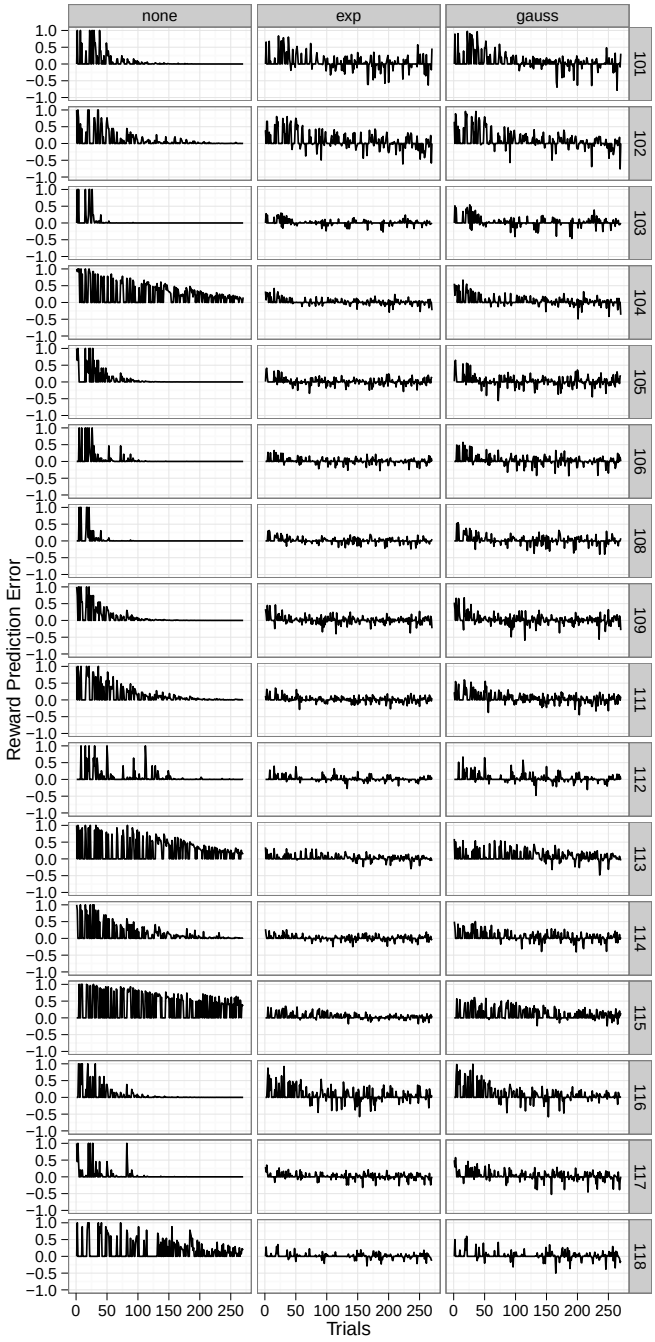

Figure 10. Reward prediction errors for each of the three models plotted for each trial in the experiment, based on the $\{1, 0\}$ coding scheme. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

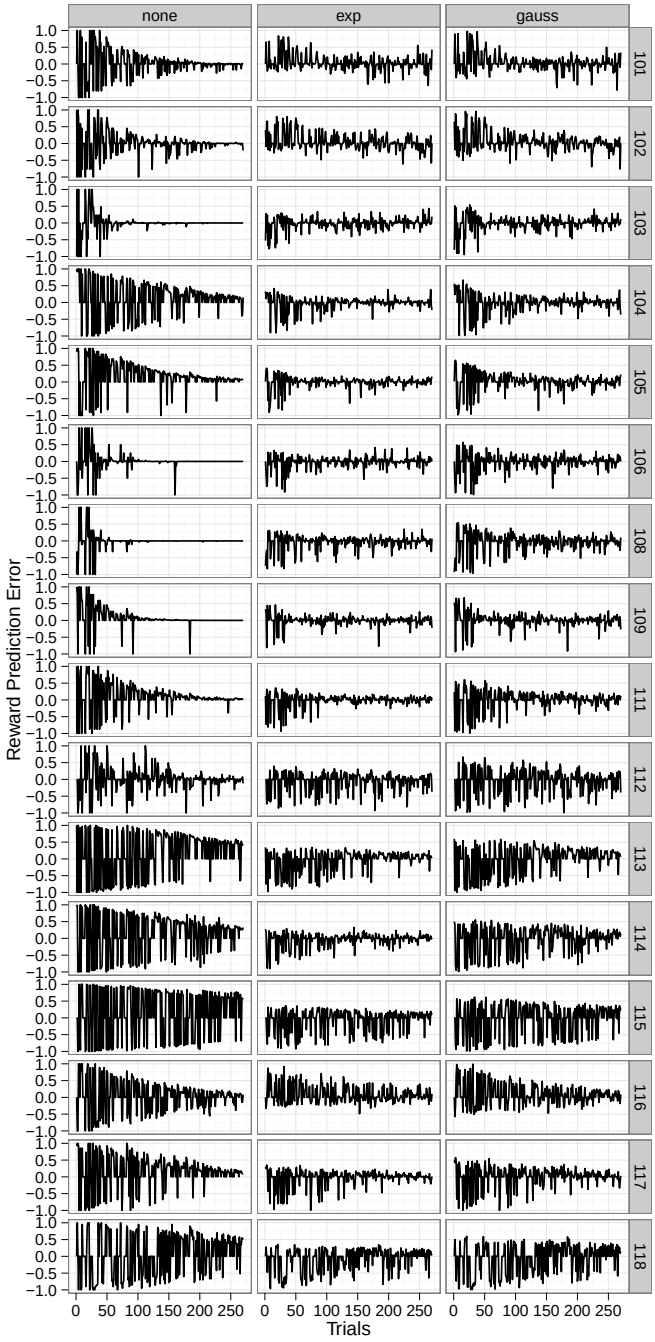

Figure 11. Reward prediction errors for each of the three models plotted for each trial in the experiment, based on the $\{1, -1\}$ coding scheme. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

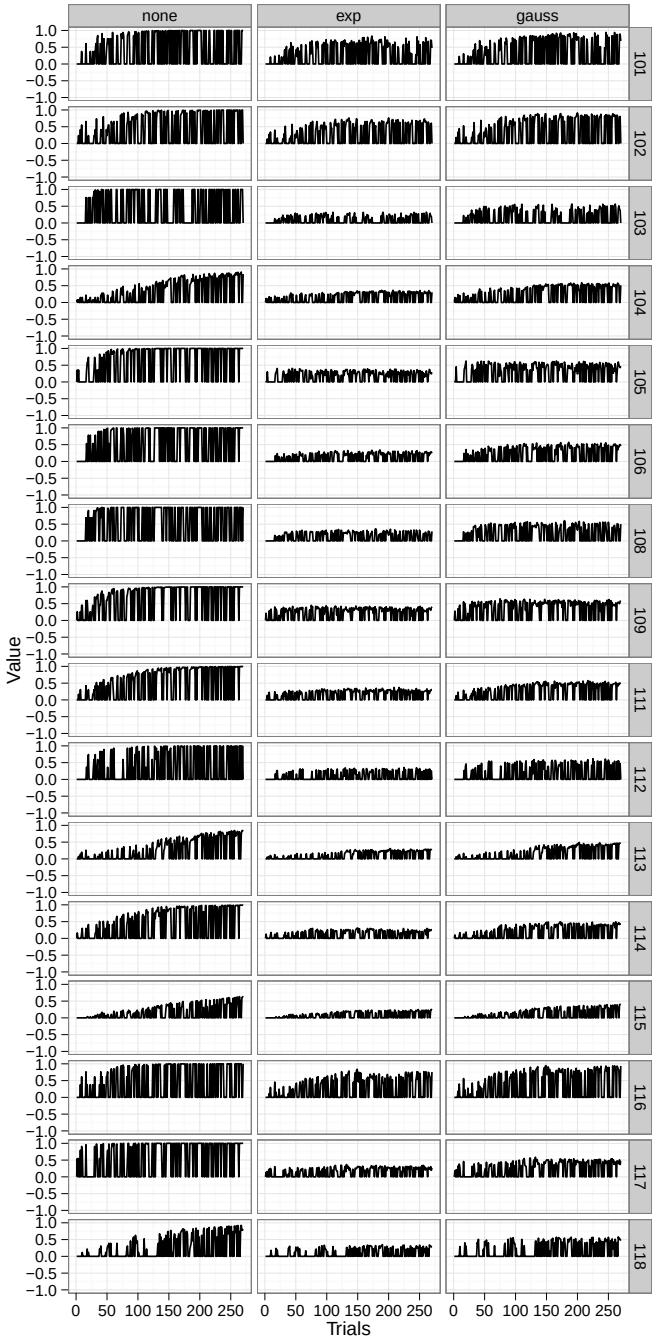

Figure 12. Value estimates for each of the three models plotted for each trial in the experiment, based on the $\{1, 0\}$ coding scheme. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

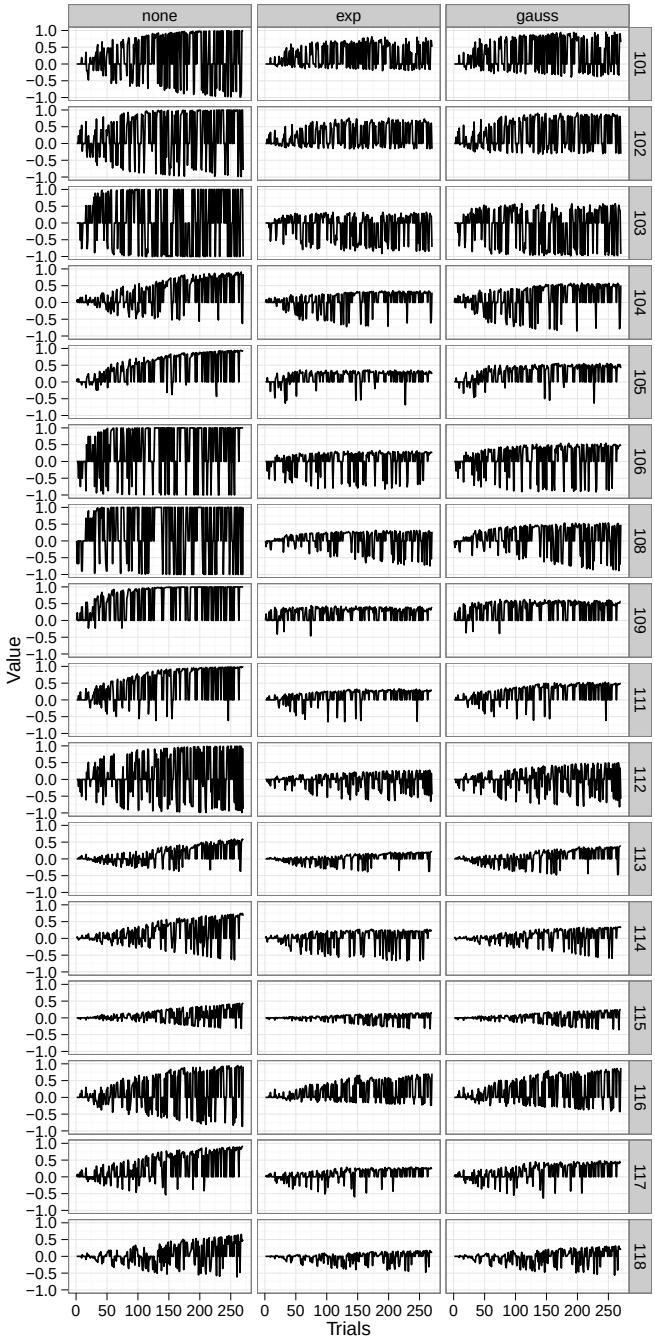

Figure 13. Value estimates for each of the three models plotted for each trial in the experiment, based on the $\{1, -1\}$ coding scheme. Each row is a single subject's data. Each column matches one of the three models, classified by their similarity metric.

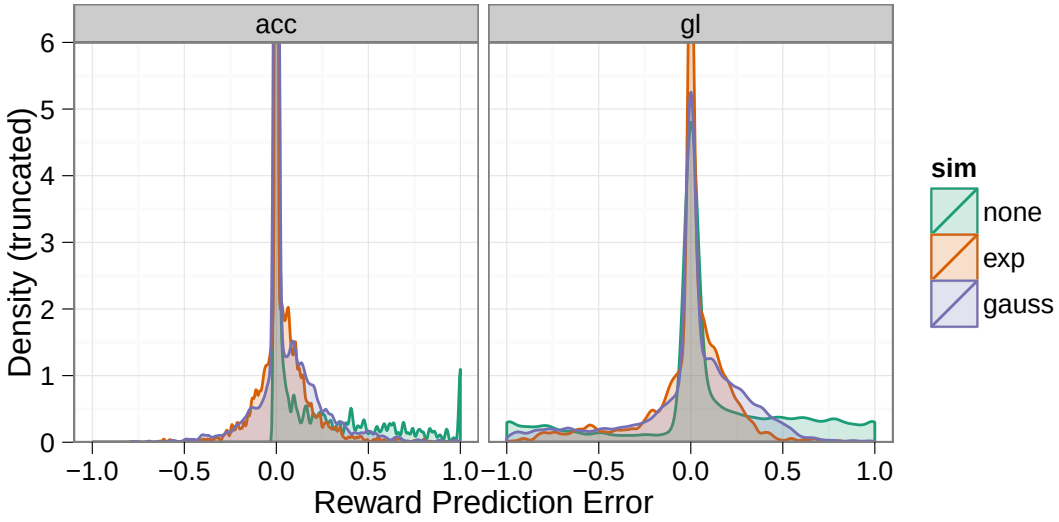

Figure 14. Density of reward prediction errors for all subjects. The y axis is truncated at 6 to allow clear visualization of non-zero values.

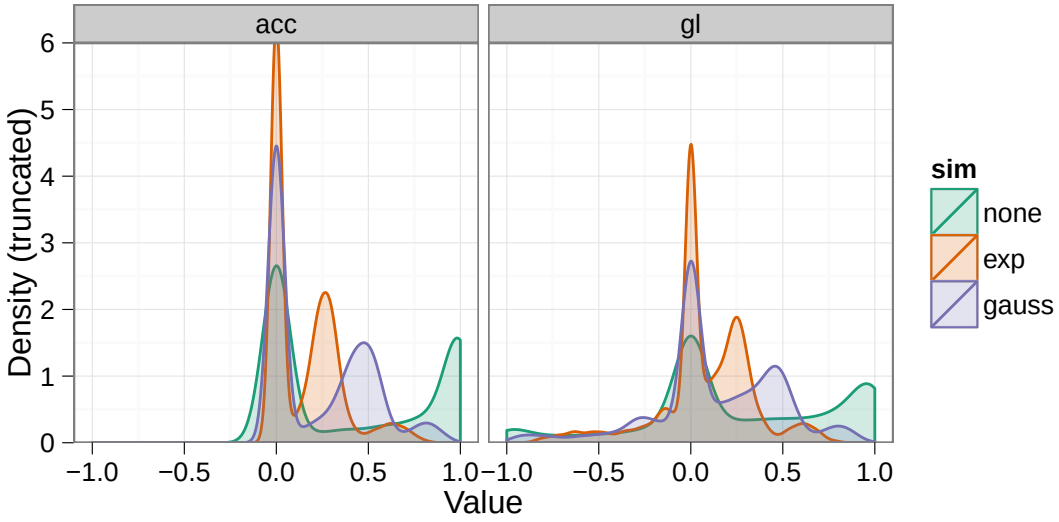

Figure 15. Density of value estimates for all subjects. The y axis is truncated at 6 to allow clear visualization of non-zero values.

References

Ashby, F. G., & Alfonso-Reese, L. (1995, Jan). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Ashby, F. G., & Ennis, J. (2006). The role of the basal ganglia in category learning. *Psychology of Learning and Motivation*, *46*, 1–36.

Ashby, F. G., & OBrien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Science*, *9*(2), 83–89.

Beierholm, U. R., Anen, C., Quartz, S., & Bossaerts, P. (2011, Jul). Separate encoding of model-based and model-free valuations in the human brain. *NeuroImage*.

Bischoff-Grethe, A., Hazeltine, E., Bergren, L., Ivry, R. B., & Grafton, S. T. (2009, Jan). The influence of feedback valence in associative learning. *Neuroimage*, *44*(1), 243–51.

Daw, N., & Courville, A. (2007). The pigeon as particle filter. *NIPS*, *20*.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011, Mar). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–15.

Delgado, M. R., Stenger, V. A., & Fiez, J. A. (2004, Sep). Motivation-dependent responses in the human caudate nucleus. *Cereb Cortex*, *14*(9), 1022–30.

Gershman, S., Pesaran, B., & Daw, N. D. (2009, Oct). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, *29*(43), 13524.

Guttman, N. (1956, Jan). Discriminability and stimulus generalization. *Journal of Experimental Psychology*.

Hornak, J., O'Doherty, J. P., Bramham, J., Rolls, E. T., Morris, R. G., Bullock, P. R., et

al. (2004, Apr). Reward-related reversal learning after surgical excisions in orbitofrontal or dorsolateral prefrontal cortex in humans. *Journal of cognitive neuroscience*, *16*(3), 463–78.

Kim, H., Shimojo, S., & O'Doherty, J. P. (2006, Jul). Is avoiding an aversive outcome rewarding? neural substrates of avoidance learning in the human brain. *PLoS Biology*, *4*(8), e233.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005, May). Distributed neural representation of expected value. *J Neurosci*, *25*(19), 4806–4812.

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, *459*(7248), 837–841.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003, Apr). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–37.

O'Doherty, J. P., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001, Jan). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci*, *4*(1), 95–102.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350.

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. J. (2007, May). Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, *27*(18), 4826.

Shepard, R. (1987, Sep). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Spiering, B. J., & Ashby, F. G. (2008, Sep). Response processes in information-integration category learning. *Neurobiol Learn Mem*, *90*(2), 330–8.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. *MIT*

Press.