Supplemental Results

The full model equations are:

Initialization: $V_i(0) = 1$ (for all actions i)

$c_i(0) = 0$ (for all actions i)

Learning: $V_i(t) = V_i(t - 1) + a * (r(t) - V_i(t – 1))$ ((for chosen action i and obtained reward r(t))

$V_j(t) = V_j(t - 1)$ (for non-chosen actions j)

$c_i(t) = 1$ (for chosen action i)

$c_j(t) = d * c_j(t - 1)$ (for non-chosen actions j)

Choice: $P_i(t) = \exp(\beta * (V_i(t - 1) + b * c_i(t - 1))) / \sum_j \exp(\beta * (V_j(t - 1) + b * c_j(t-1)))$
where t indexes trials

There are, in total, four free parameters. The learning rate a controls how sharply the model updates the expectation $V_i(t - 1)$ toward the observed reward r(t). The perseveration bias weight b controls the strength of the model's tendency to choose (for $b > 0$) or avoid (for $b < 0$) recently chosen options; the decay factor d controls over how many trials this bias persists after a choice. Finally, the softmax inverse temperature β controls the randomness of the choices: for large β, the model is more likely to choose the action believed to have the maximum value. Thus, larger β will result in a lower randomness of choices

We fit the model parameters to behavior in two ways: individually (one parameter set per subject) and groupwise (one parameter set each for all Learners and all Non-learners; we refer to this as a “fixed effects” model since it treats each parameter as fixed within the group).

As we have noted previously (Daw et al., 2006) individual parametric fits in tasks and models of this sort tend to be noisy, probably owing in part to the fact that maximum likelihood estimates have no regularization; in our experience, regularization of the parameter estimates over the population tends therefore to improve a model's subsequent fit to fMRI data. Fitting the parameters as fixed within the groups is a simple form of regularization, and we therefore (following previous work; Daw et al., 2006; O'Doherty et al. 2004) used the fixed effects estimates to generate regressors for fMRI. For completeness, these parameters are shown in Supplemental Table 1S, together with confidence intervals derived from an asymptotic covariance estimator (inverse Hessian of log data likelihood).

To investigate behavioral differences between groups and across individuals, we use the individual fits, whose summary statistics are shown in Supplemental Table 2S, instead of the fixed effects fits. A further complication with using TD fits to compare groups is that the parametric estimates are not independent but instead tend to covary over subjects. In particular, because each trial's feedback is multiplied by the learning rate to compute the value, and this value itself is multiplied by the softmax temperature to compute logit choice probability, these two parameters tend to be inversely coupled. Therefore, as seen in the tables, the parameters viewed separately can have improbable means and large estimation errors (and neither differed significantly between groups; Wilcoxon rank sum test, $p > .15$). However, their product β*a tends to be more reliably estimated. Because, multiplied together, these parameters control how strongly a particular reward impacts subsequent choice preferences (i.e., logit choice probability), their product is also a more appropriate measure of trial-to-trial sensitivity to reward than either parameter individually. It is also comparable to the weight on the most recent reward in a logistic regression analysis of choices (e.g., Lau & Glimcher, 2005, c.f. Lohrenz et al. 2007).

We therefore tested whether these trial-to-trial reward sensitivity estimates, β*a from the individual parametric fits, reflected the observed behavioral differences between the groups. Indeed, learners were significantly more sensitive to rewards on this metric than Non-learners (Wilcoxon rank sum test, $p < 0.0005$). We additionally examined whether these estimates correlated across subjects with the measure of behavioral performance used to classify Learners and Non-learners (i.e., the number of choices on the high probability decks, 75% and 60%, in the last 40 trials of the task). Indeed, as shown in Supplemental Figure 1S, the reward sensitivity estimate correlated with the learning criterion (linear regression, $r^2=0.43$, $p < 0.0005$; this correlation and also the between-group comparison discussed above remain significant when the outlier at the top right corner of the figure is eliminated).

Of the remaining parameters, the perseveration bias decay d did not differ significantly between subjects (Wilcoxon rank sum test, $p = 1$), while the perseveration bias weight b did ($p < .02$; in fact, since b is also multiplied by β in the softmax, the parameters are similarly coupled and the difference was more reliably detected in the product β*b; $p < .002$). This result indicates that behaviorally Learners and Non-learners differ not just in their sensitivity to rewards, but also in their tendency to revisit or avoid previously chosen options regardless of their reward history. The fit parameters suggest a tendency of Learners to stick with previous choices ($b > 0$); Non-learners the opposite (This may simply reflect the former group's defining strategy of finding the best option and then sticking with it).

Finally, we investigated whether we could detect sensitivity to rewards even in Non-learners, by re-fitting the model to this group with the learning rate(s) restricted to zero. With this restriction, rewards cannot impact choices, which can instead only be explained using choice autocorrelation (through parameters *b* and *d*). We fit this restricted model in two ways, with the remaining free parameters either fit individually or as fixed effects over the Non-learners group. In both cases, the null hypothesis of zero learning rate(s) was rejected compared to the corresponding full model with nonzero learning rate(s) (likelihood ratio test, fixed effects: 1 d.f., p<.005, individual fits summed over group: 12 d.f., p<0.000000001). These results add additional support to the conclusion that Non-learners were sensitive to the rewards.

References

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans, Nature 441:876-879.

Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys, J Exp Anal Behav 84:555-579.

Lohrenz T, McCabe K, Camerer CF, Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task, Proc Natl Acad Sci USA 104:9493:9498.

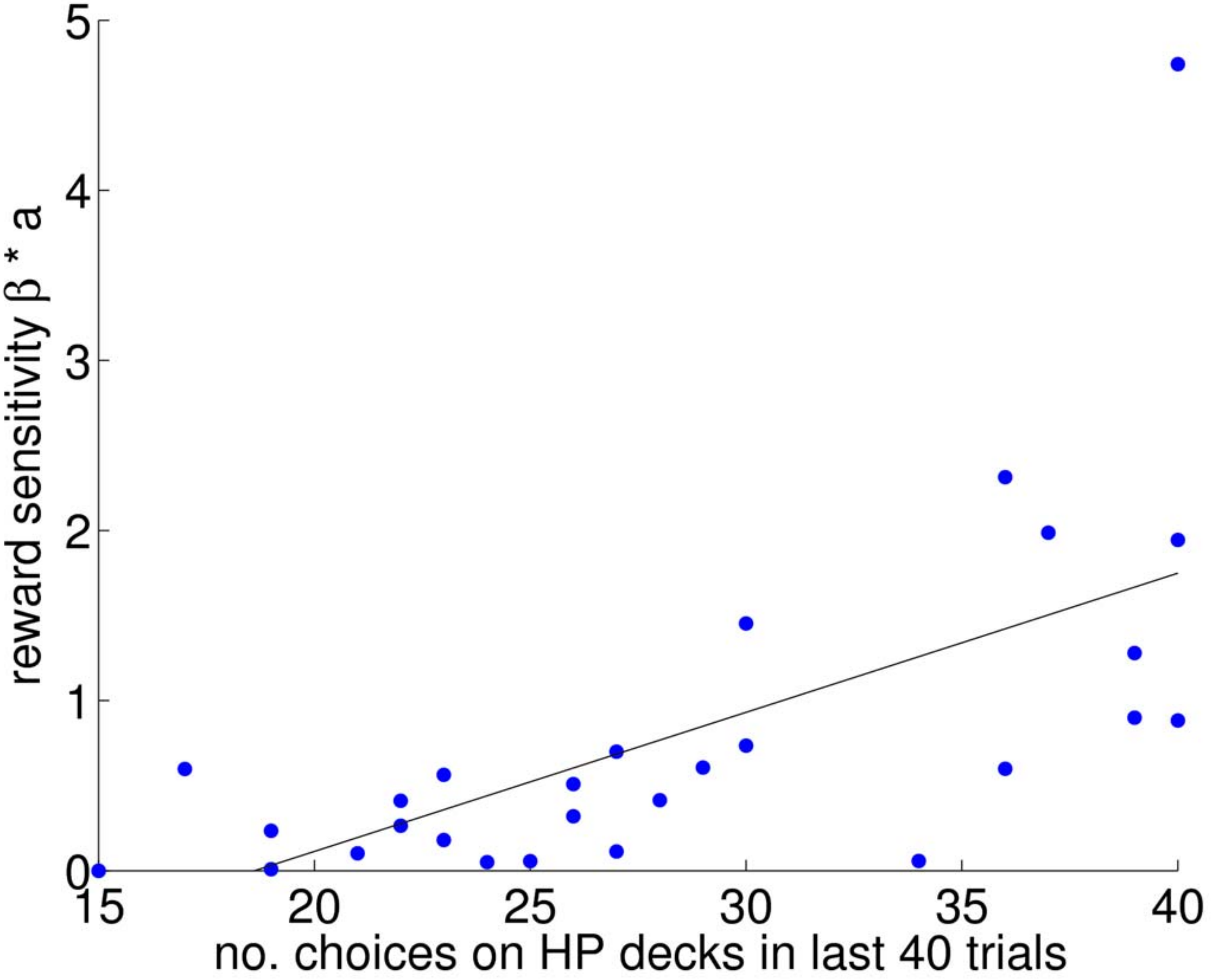
O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning, Science 304:452-454.

	Learners	Non-learners
learning rate *a*	.33 +/- .03	.37 +/- .48
softmax inv. temp. β	2.2 +/- .15	.49 +/- .26
choice weighting *b*	1.1 +/- .10	-1.7 +/- .92
choice decay multiplier *d*	.096 +/- 4.2e-3	.70 +/- .05
$\beta * a$	.75 +/- .055	.18 +/- .23

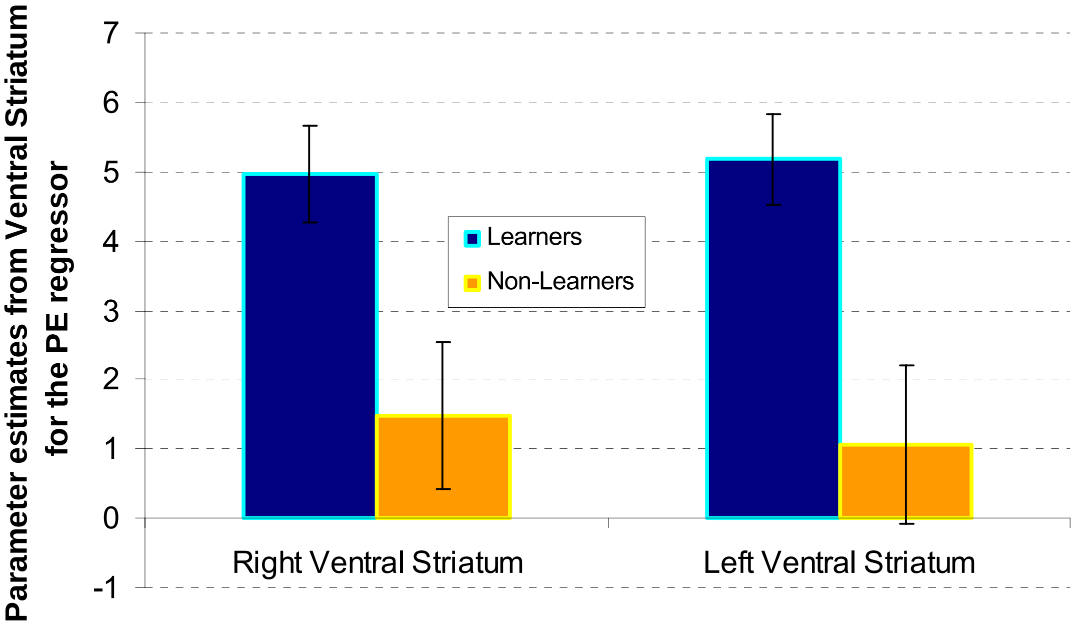
Supplemental Table 1S: Parameters fit as fixed effects to group behavior (shown +/- one standard deviation from asymptotic covariance estimator). Also shown are the moments for the product of the softmax temperature and learning rate, taking into account their covariation.

	Learners	Non-learners
learning rate *a*	.28 +/- .05	.25 +/- .11
softmax inv. temp. β	7.6 +/- 2.4	1.0e+5 +/- 7.0e+4
choice weighting *b*	.12 +/- .14	-10 +/- 5.0
choice decay multiplier *d*	.52 +/- .10	.55 +/- .10
$\beta * a$	1.2 +/- .28	.21 +/- .06

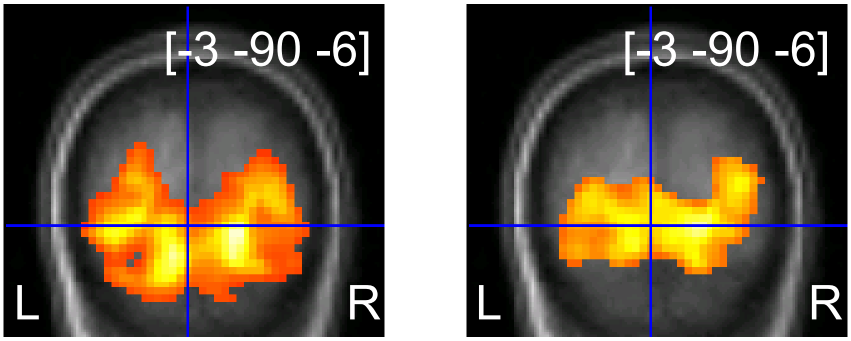
Supplemental Table 2S: Mean (+/- 1 SEM) over subjects of parametric fits to individual behavior. Also shown are the moments for the product of the softmax temperature and learning rate.



Supplemental Figure 1S: Scatter plot comparing two measures of behavior over subjects: the number of choices on the high probability (HP) decks (75% and 60%) in the last 40 trials of the task vs. the product of learning rate and softmax temperature (a measure of trial-to-trial sensitivity to reward) from individual parametric fits of the TD model.



Supplemental Figure 2S: Parameter estimates from left and right ventral striatum for the PE regressor reported separately for the Learner and Non-learner groups. The parameter estimates were extracted separately from the left and right striatum at the MNI co-ordinates of (-9, 12, -12) and (+9 ,12,-12) respectively.



Supplemental Figure 3S: Activity in visual cortical areas related to the trial onset during task performance in both Learner and Non-learner groups. Highly significant responses were found in this region in both groups (A) Learners; (B) Non-learners, at p<0.0001 (uncorrected). Furthermore, no significant differences in activity were found between the groups in this area in a direct contrast at p<0.001 uncorrected. These findings support the claim that subjects in both groups were processing the visual components of the task and argues against the possibility that Non-learners were simply disengaged from the task.

Brain Region	Laterality	X	Y	Z	Z-Score
Cingulate Gyrus	L	-18	-24	39	4.31
Parahippocampus	L	-33	-18	-30	4.21
Anterior Cingulate Gyrus	R	18	30	24	3.97
Medial Orbital Gyrus	L	-3	60	-15	3.79
Cingulate Gyrus	R	12	-3	36	3.66
Putamen	R	30	-9	9	3.65
Cingulate Gyrus	R	18	-30	39	3.62
Hippocampus	R	33	-21	-12	3.52
Insula	R	36	3	-12	3.48
Superior Frontal Gyrus	L	-12	63	-6	3.4
Inferior Frontal Gyrus	L	-48	3	24	3.34
Inferior Temporal Gyrus	L	-27	-9	-30	3.32
Medial Frontal Gyrus	L/R	0	45	24	3.27

All peaks are of clusters with an extent of min. 5 voxels
All peaks survive a significance threshold of p < 0.001

Supplemental Table 3S: Regions outside of our striatal regions of interest correlating with prediction error in the contrast of Learners minus Non-learners (p<0.001, uncorrected)